Dissecting Transcriptional Regulatory Networks with Systems Biology Approaches

Xiang Zhou

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

under the Executive Committee

of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2011

ABSTRACT

Dissecting Transcriptional Regulatory Networks with Systems Biology Approaches

Xiang Zhou

In the past decade, technologies such as the DNA microarray and ChIP-on-chip have generated a large amount of high-throughput data for biologists. Although these data has provided us systems-level information about gene regulation, a major challenge in systems biology is to derive methodologies that will infer the underlying dynamics and mechanisms of gene regulation. This thesis research is focused on understanding these mechanisms of transcriptional regulation using systems biology approaches. Transcription regulatory networks play an important role in mediating external stimuli and coordinating responses to changing environments. Different methods that infer regulatory interactions directly from microarray data have been developed in the recent past. However, the implicit assumption in these methods — that the transcription factor (TF) mRNA expression can be used as a proxy of its activity at protein level — is not always correct, due to post-transcriptional and post-translational modifications of TFs. In this study, a method named iARACNe was developed. It uses the inferred TF activities to estimate the regulatory activity between TFs and their targets. The study demonstrated that the accuracy of the inferred networks using this method was greatly improved. Two additional methods, OmniMiner and coEDGi, which allow a better understanding of the physical interactions between TFs and target genes, were developed in this thesis research. OmniMiner detects and predicts the potential binding sites for the TFs of interest, while coEDGi enables identification of common enhancers upstream of co-regulated genes. Compared to other approaches which only allow isolated analyses, the systems biology approaches developed in this

research provide an opportunity for biologists to study transcriptional regulations from both functional genomics and regulatory sequence perspectives simultaneously.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

Five years may seem long to someone, but when you are doing something you like, time flies like an arrow.

First of all, I would like to thank my advisor and mentor, Professor Andrea Califano. Thank you for allowing me studying and growing in your lab. It was an amazing journey. I enjoyed every moment of it and learned a lot of from it. You are a great PI and will always be my role model.

Second, I would like to thank my committee members, Prof. Richard Mann, Prof. Harmen Bussemaker, Prof. Gustavo Stolovitzky and Prof. Raul Rabadan. Thank you for spending your valuable time serving on my committee and giving me advices on my thesis projects.

Third, I would like to thank all my collaborators, Celine Lefebvre, Mariano Alvarez, Mukesh Bansal and Jorida Coku on iARACNe project, Pavel Sumazin and Presha Rajbhandari on OmniMiner project, and Prof. Richard Mann, Dr. Matthew Slattery and Dr. Matt Giorgianni on coEDGi project. Without you contributions and suggestions I couldn't go this far.

Forth, I would like to thank all the members in Prof. Califano's lab. You are like a family to me. I am honored and happy being a member of this group. You guys rock.

Last, I would like to especially thank my wife and my parents. My beautiful wife, Yang, gave me tremendous supports over all these years and just gave birth to our lovely son, Ethan, this January. I love you and will always be. And to my mom and dad, thank you for encouraging me chasing my dream. I love you.

Chapter 1

The general focus of my research is to improve our understanding of transcriptional regulation mechanisms. More specifically, I am trying to understand how transcriptional processes may elucidate transformation between normal and disease-related condition in the cell. In my research, I studied this question both from a functional genomics perspective, using gene expression profile data, and from a regulatory sequence perspective, based on DNA binding signatures. I showed that it is possible to develop systems biology approaches to help identify key players in transcriptional regulation processes as well as their interplay in regulating specific phenotypes.

In Chapter 2, I will discuss the algorithm iARACNe which enables construction of robust transcriptional regulatory networks. This algorithm is an extension of the well-established ARACNe algorithm (Margolin, Nemenman et al. 2006) to address limitations of network inference methods that use transcription factor (TF) mRNA expression as a proxy of its activity at the protein level. For TFs that are significantly post-transcriptionally regulated, this approximation is not appropriate and may severely degrade performance of network inference methods. We propose to address this challenge by first inferring and then plugging the TF protein activity in the network reconstruction process. Using ChIP-on-chip data and expression profiles following TF silencing, we show that networks elucidated by inferred TF protein activity was more accurate than when TF mRNA expression alone was used. This new network

construction approach provides an opportunity improve our understanding of functional relationships between TFs and their targets in a specific cellular context.

In Chapter 3, the OmniMiner motif discovery algorithm will be described. This algorithm integrates both functional and sequence information to predict the potential DNA binding motif of a given TF. Functionally, the ARACNe algorithm is used to identify candidate direct targets of the TF as an input to the algorithm, under the assumption that their promoters will be significantly more enriched in the TF DNA binding sites than genes that are simply co-expressed with TF. Using sequences from these targets greatly improved the signal to noise ratio and increased the probability of identifying accurate DNA binding motifs. We combined this approach with integration of alignment-based and pattern-discovery-based (alignment-free) information to further increase our ability to identify high-probability TF binding regions and sites. OmniMiner significantly outperformed existing DNA binding motif discovery approaches and is unique in that it may be applied to predict binding motifs for TFs within specific cellular contexts. This work has been published by PLoS One, in 2010, where it was the highest downloaded bioinformatics paper for several months (Zhou, Sumazin et al.).

In Chapter 4, coEDGi, an enhancer discovery algorithm, will be described. This algorithm is an extension of the previously published EDGi algorithm (Sosinsky, Honig et al. 2007). EDGi shows that use of non-alignment based pattern discovery methods can significantly improve our ability to identify distant enhancer modules in higher

eukaryotes. However, EDGi is only able to analyze one gene at a time and does not take advantage of functional data related to gene regulation. coEDGi extends EDGI by integrating gene co-regulation information into the enhancer discovery process, which allows discovering common enhancers upstream of co-regulated genes. Use of co-regulation data significantly increases the resolution and predictive power of enhancer inference, when compared to use of EDGi alone.

Chapter 2

Reconstruct transcriptional regulatory network with inferred transcription factor protein activities

2.1 Introduction

"No man is an island, entire of itself; every man is a piece of the continent, a part of the main;…" Meditation XVII, No Man is an Island. John Donne (1572-1631)

Much like in the mediation of John Donne that no man is an island, no gene is an island unto itself. Further, just as we are not sole actors in life, but connected "to this main" in all our actions, any biological function is rarely performed by a single gene (Barabasi and Oltvai 2004). Thanks to the abundance of high-throughput data in the past ten years, we are now able to study the transcriptional regulation of genes from a global view. I am interested in developing systems biology approaches to improve our understanding of transcriptional regulatory mechanisms. In my opinion, systems biology approaches are not merely the mathematic equations that mimic the cell behaviors, but potential models or strategies that integrate different analyses from different data sources, and, by doing so, help to identify key players in transcriptional processes. My research applied this concept to study transcriptional regulatory networks from a functional genomics perspective.

Microarray high-throughput data have been widely used to reverse-engineer transcriptional regulatory networks for many years (D'Haeseleer, Wen et al. 1999; Butte and Kohane 2000; D'Haeseleer, Liang et al. 2000; Friedman, Linial et al. 2000; Hartemink, Gifford et al. 2002; Gardner, di Bernardo et al. 2003; Imoto, Higuchi et al. 2003; Basso, Margolin et al. 2005; di Bernardo, Thompson et al. 2005; Margolin, Nemenman et al. 2006). Different models were proposed through this body of work, and can be categorized into four groups: 1) coexpression-based networks, 2) Bayesian networks, 3) Ordinary differential equations, and 4) information theory approaches. A brief description of the advantages and disadvantages of these four models is described immediately below.

Strictly speaking, the coexpression-based approach is not a network inference algorithm. This type of approach assumes that coexpressed genes are likely to be functionally related. Therefore, genes can be measured by a distance metric, such as correlation coefficient. Based on their distances, genes can be clustered into different groups. But in reality, coexpressed genes do not guarantee functional similarity. In addition, we cannot tell the causal relationship between genes nor determine whether genes are directly or indirectly connected. Despite its limitations, however, coexpression-based clustering is still one the most popular approaches to analyze microarray data.

Compared to the coexpression-based approach, the Bayesian network (BN) is a more sophisticated method. It is a graphic model that measures the probabilistic

relationships among a set of random variables X_i, where i =1… n. And those relationships are represented as a joint probability distribution, $P(X_1, X_2, \ldots, X_n)$. BN is a directed acyclic graph (DAG) in which we assume that each variable is independent of its non-descendants. Therefore, the joint probability distribution can be represented as **Equation 2-1**:

$$P(X_i, \ldots X_n) = \prod_{i=1}^{N} P\left(X_i = x_i \middle| X_j = x_j, \ldots, X_{j+k} = x_{j+k}\right)$$

Equation 2-1

The Bayesian network model identifies a DAG that best represents the expression data (D). Therefore, a scoring function is needed to evaluate each graph (G). That being said, it is impossible to try out all gene combinations to find out the best G. Heuristic search methods, such as greedy hill-climbing approach and the Markov Chain Monte Carlo method, among others, are used during model prediction. BN has a number of features that make it an attractive approach, such as handling incomplete data well and avoiding over-fitting the model with the training data. But BNs also have several limitations. For example, because of the DAG structure, BN doesn't allow feedback loops, even though feedback loops constantly exist in real networks. Dynamic BN partially solved this problem by separating input nodes from output nodes. For instance, genes were represented by both their parents (regulators) and targets (children) (Perrin, Ralaivola et al. 2003; Yu, Smith et al. 2004). In addition, the probabilistic dependence between genes does not guarantee causal relationships and we could not tell whether the interaction is direct or indirect. Despite these limitations, BN remains one of most popular models used to reconstruct regulatory networks (Friedman, Linial et al. 2000; Hartemink, Gifford et al. 2002; Perrin, Ralaivola et al. 2003).

Transcriptional network can also be inferred by using ordinary differential equations (ODEs), which describe gene expression changes as a function of other genes and external perturbations (**Equation 2-2**),

$$x_i(t) = f_i(x_1, \ldots, x_1, u, \theta_i)$$

Equation 2-2

where "θ_i" is a set of parameters describing the interactions among genes, $x_i(t)$ is the expression of gene (i) at time (t), along with external perturbation (u) to the system. Network inference is described as the identification of function (f_i) and estimation of the unknown parameters (θ_i). An advantage of ODE is that once the parameters θ_i for all i are known, the behavior of the network can be quantitatively predicted under different conditions. But because of the large requirements of data inputs for ODEs, this approach was mainly applied to a relatively small network. Current applications that use ODEs are the Network Identification by multiple Regression (NIR) (Gardner, di Bernardo et al. 2003), Microarray Network Identification (MNI) (di Bernardo, Thompson et al. 2005), Time Series. Network Identification (TSNI) (Bansal and di Bernardo 2007).

In the current study, an information theory based approach was used. Mutual information (MI) was applied to detect the pairwise dependencies between genes applying **Equation 2-3**,

$$MI_{ij} = H_i + H_j - H_{ij}$$

Equation 2-3

where H, the entropy, is defined as **Equation 2-4:**

$$H_i = -\sum_{k=1}^{n} p(x_k)\log\left(p(x_k)\right)$$

Equation 2-4

The higher MI indicates that the two genes are not randomly associated. If MI is zero, it suggests that two genes are statistically independent of each other. The early implementation of using information theory to reconstruct networks was proposed by Butte and Kohane as a "relevance network" (Butte and Kohane 2000). However, the interactions predicted by relevance network contained a lot of false positives. In addition, the edges in the network didn't tell us the direction of the interactions. ARACNe developed by Margolin et al (Margolin, Nemenman et al. 2006) was also based on the information theory, but solved the problem with relevance network. Because ARACNe was used to reconstruct transcriptional regulatory network, the interactions in the network were between transcription factors and their potential target genes. Therefore, in ARACNe, a pre-defined TF list was required and MIs were computed between TFs from the list and all other genes in the dataset. This strategy enabled identification of the causal relationship between TF and the target gene. Another improvement in ARACNe was that Data Processing Inequality (DPI) was applied to detect direct interactions. The rationale was that if both (x, y) and (y, z) are directly interacting with each other and (x, z) is indirectly connected through y, the MI between x and z should be no larger than either MI between x and y, or y and z, $MI_{x,z} <= \min(MI_{x,y}, MI_{y,z})$. By removing the weakest link in the triplet, ARACNe dramatically reduced the false positive rate and inferred a high percentage of direct interactions. For a more detailed description of ARACNe, please

refer to: (Basso, Margolin et al. 2005; Margolin, Nemenman et al. 2006; Margolin, Wang et al. 2006); and Material and Method section.

ARACNe has been shown to work well for reconstructing large-scale regulatory networks. Basso *et al* (Basso, Margolin et al. 2005) applied ARACNe to reverse engineer the regulatory network for human B-cells and showed that MYC targets inferred from the network were highly accurate and directly regulated by MYC. In 2010, ARACNe was used to build the transcriptional network for mesenchymal transformation of brain tumors (Carro, Lim et al.), and Lefebvre *et al.* built human B-cell interactome with ARACNe as well as identified the master regulators of proliferation in germinal centers (Lefebvre, Rajbhandari et al.). Their work suggested that ARACNe algorithm is a reliable method in reconstructing regulatory networks.

But ARACNe faces the same limitation as all other network construction methods that use mRNA expression as a proxy to protein activity. As there is no genome-wide protein activity measurement approach, almost all network reconstruction algorithms use the gene mRNA expression level as a proxy to its protein activity. But in reality, there are multiple procedures between mRNA and protein, such as mRNA slicing, mRNA stability, *etc*. This process is called post-transcriptional regulation. Studies have shown that a large number of transcription factors were post-transcriptionally regulated (Day and Tuite 1998; Lee, Colinas et al. 2006; Chen and Rajewsky 2007; Fu, Drinnenberg et al. 2007; Filipowicz, Bhattacharyya et al. 2008). For example, microRNAs, approximately 21-nucleotide-long non-coding RNAs, are one of the key regulators for gene post-transcriptional regulation (Filipowicz, Bhattacharyya et al. 2008). For TFs that

have been post-transcriptionally regulated, mRNA expression levels were no longer an accurate estimation for the protein activities. It has been shown that TF activities can be quantitatively predicted by a multivariate regression model on the expression data (Bussemaker, Li et al. 2001; Keles, van der Laan et al. 2002; Wang, Cherry et al. 2002; Conlon, Liu et al. 2003). Gao *et al* integrated ChIP data and mRNA expression data in their MA-Networker model which inferred the activity of transcription factors (Gao, Foat *et al*. 2004). And very recently, Youn et al. developed a probabilistic model that also integrated the location and expression data for network construction (Youn, Reiss et al.). However the binding information from either ChIP data or DNA motif data is a prerequisite for these methods. Although relatively easy to obtain in yeast genome, they are currently limited in the human genome. Therefore, there is a need to develop a new approach that enables us to measure TF protein activity in higher species, such as humans, from their mRNA expressions. We introduce a new algorithm, iARACNE (iterative ARACNE), that addresses this issue by first predicting TF protein activity from a first round of reverse engineering and then using the inferred activity of the TF in place of its mRNA expression in one or more subsequent target-inference rounds. This significantly improves the accuracy of the inferred transcriptional network. Specifically, first, the standard version of ARACNE is used to infer an initial complement of targets of each TF of interest (i.e. its regulon). While these regulons may have higher false positive rates, the assumption is that their global activity still provides a better estimate of the TF protein activity. The virtual protein activity is computed by measuring the enrichment score (ES) of each TF's regulon in the expression of each sample using Gene Set Enrichment Analysis (GSEA). The computed ES is then used instead of the TF mRNA

expression in a second run of ARACNE. This can be repeated, if necessary, until convergence (i.e. no additional changes to the regulon composition). Normalization is not necessary because regulon size is constant across samples for each TF. iARACNE was tested on multiple cancer data sets, including B-cell lymphomas, high grade glioblastoma and germ cell tumors. ARACNE and iARACNE were compared by computing: (1) network likelihood ratio (ratio of probabilities of true positive and false positive interactions, by comparing all inferred interactions to known and randomly generated interactions); (2) regulon enrichment analysis from experimental assays, including ChIP and TF silencing. Likelihood ratios showed almost a two-fold improvement in iARACNE compared to ARACNE due to the significant decrease of false positives. Regulon enrichment analysis was performed to evaluate the recalls for STAT3 and BCL6 inferred regulons of experimentally validated targets. In both cases, iARACNE showed a two- to three-fold improvement over ARACNE. In summary, the new algorithm improves ARACNE's performance by inferring a TF's activity as well as its targets. There are two novelties in our method. First, the method can effectively estimate TFs' activity without requiring ChIP-on-chip data and can be further improved when it is available. Second, post-transcriptional regulations were taken into consideration during network reconstruction and this inclusion greatly improved the accuracy of the inferred TF-target interactions in the network.

2.2 Materials and Methods

2.2.1 Data sources

In this study, two cancers were focused on: leukemia and glioblastoma multiforme (GBM). The B-cell leukemia data contained nine phenotypes, including both normal and disease conditions, such as diffused large B cell lymphoma (DLBCL), follicular lymphoma (FL), Chronic Lymphocytic Leukemia (CLL) and *etc*. For detailed information of the phenotype data please refer to Lefevre *et al* (Lefebvre, Rajbhandari et al.). The GBM data was obtained from *The Cancer Genome Atlas* (TCGA) (2008) using the Affymetrix HT-HG-U133A platform, in which there were 319 patient samples in the data set.

2.2.2 Quality control procedure

The experiment was started by first checking the quality of the expression data with principle component analysis (PCA). As shown in **Figure 2-1**, samples from the same leukemia subtypes were classified together. Because GBM data was collected from different patients, there was no need to do PCA on that data set. The results showed that the data was of high quality.

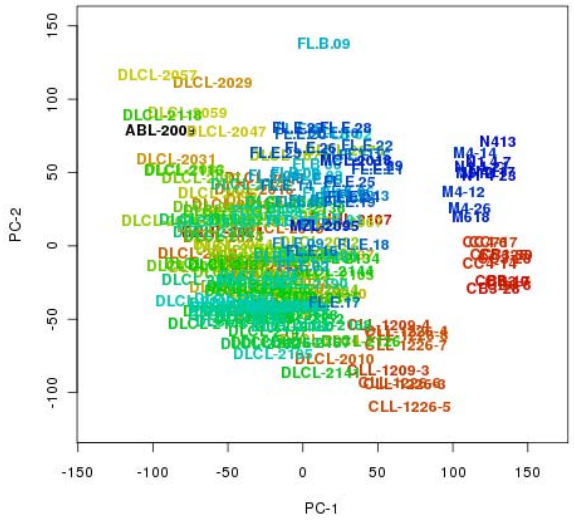


Figure 2-1 PCA analysis of B-cell expression data

2.2.3 Normalizing expression data with z-score transformation

Within iARACNe, gene set enrichment analysis was used to estimate the enrichment score for TFs. To compute a transcription factor's enrichment score under each condition, to the first step is to normalize a gene's expressions across all samples so that they can be compared at a same scale. Then z-score transformation (Cheadle, Cho-Chung et al. 2003) was then applied to normalize gene expression across all sample conditions. Formula 1-1 was used to compute gene expression z-scores,

$$Zscore_{si} = \frac{(I_{si} - u_{si,\ldots,sn})}{SD_{si,\ldots,sn}}$$

Equation 2-5

where I_{si} was the intensity at sample I, $u_{si,\ldots,usn}$ was the mean of the intensities across all samples and $SD_{si,\ldots,sn}$ was the standard deviation of the intensities across all samples. The z-score transformed expression file was used as the input for iARACNe.

2.2.4 Gene set enrichment analysis

GSEA (Gene Set Enrichment Analysis) (Subramanian, Tamayo et al. 2005), developed at BROAD institute, is a method that estimates whether a set selected genes are statistically significantly enriched in one biological condition (a.k.a. sample) versus another. In this study, the initial definition of GSEA was modified to fit our model design. Instead of comparing two extreme biological conditions, the selected sample was compared to the mean of all samples. The GSEA procedures are as follows: a) for each of the N genes (N: the total number of genes in the selected sample), order them by their normalized z-scores from the most positive to the most negative values, donated by R. b) Identify hits independently for the positive gene set S+ (genes that were activated by TF) in R and the negative gene set S- (genes that were repressed by TF) in –R, in which –R is the inversed ranking of R with the inverted z-scores. c) Combine R and –R and re-order the z-scores by keeping the hits for both S+ and S-, donated as Rc. d) Compute a running score by walking down the combined ranking Rc. For a given position *i* in Rc, get P_{hit} or P_{miss} with Equation 2 A and B. e) The enrichment score (ES) is the maximum derivation from zero of P_{hit} - P_{miss}.

$$P_{hit}(S,i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{\sum_{g_j \in S} |r_j|^p}$$

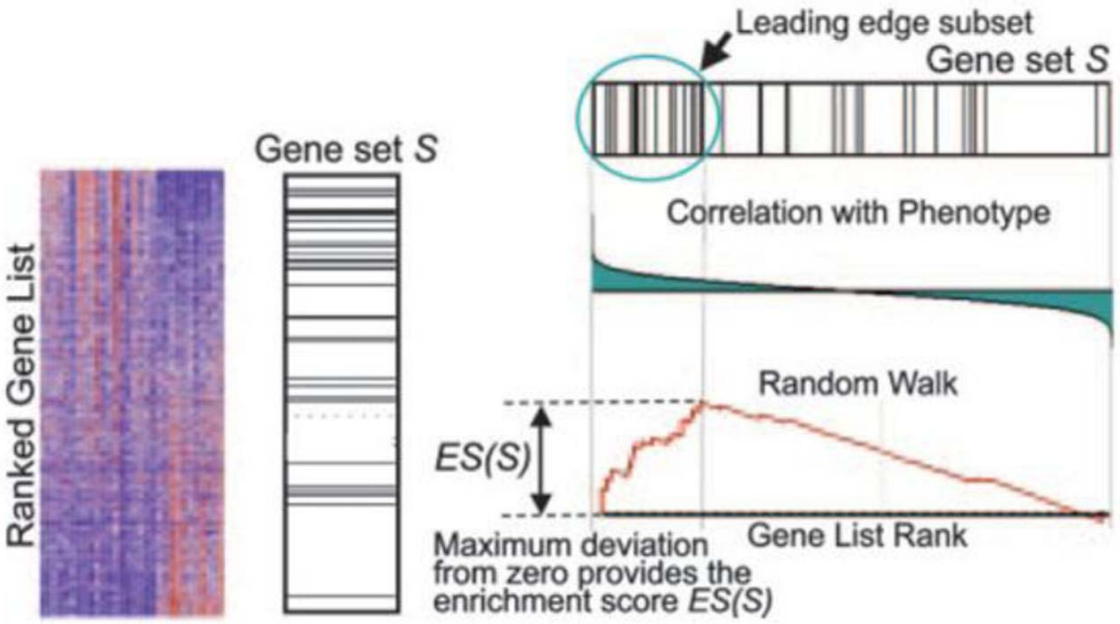
Equation 2-6

$$P_{miss}(S,i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(2N - S)}$$

Equation 2-7

(S is the combined total number of genes in S+ and S-)

One thousand permutations were done by shuffling the gene labels in Rc and repeating steps to compute the ES null distribution. Normalized enrichment score (NES) is the ES of S divided by the mean of random ESs. **Figure 2-2** is for demonstration purposes only.



GSEA demonstration (Subramanian, Tamayo et al. 2005)

Figure 2-2. Demonstration of GSEA procedure

2.2.5 iARACNe procedure

iARACNe was performed in the following steps.

1) The initial RN (a.k.a. R1) was constructed with a z-score transformed mRNA expression using the standard ARACNe algorithm (See **Materials and Methods** section for details). In the R1 network, each transcription factor had a set of predicted targets (a.k.a. TF regulon) (Margolin, Wang et al. 2006).
2) All TFs that regulated at least twenty targets were selected. Twenty is the minimal number of targets required to calculate a statistically reliable enrichment score for TF. For each sample, all genes were ranked based on their normalized z-scores from high to low and the ranked genes used as the reference list in GSEA analysis. The enrichment score was calculated by using the predicted TF regulon and then normalized. The normalized enrichment score (NES) of TF was considered as the initial TF protein activity. For all selected TFs, a NES was assigned to each sample condition.
3) The newly calculated TF protein activity (NES) was used to compute the mutation information between TF and other genes. By applying this approach, the actual relationship was measured between TF protein and the target gene mRNA expression. Standard ARACNe was applied to construct a new transcriptional regulatory network (a.k.a R2) based on the new mutation information computed between TF protein activity and gene's mRNA expression. A return to step 2 was done by using regulons from R2 and iteratively building another RN, based on R2. The third network is called R3 accordingly. More RNs were constructed until converge.

2.2.6 Network likelihood ratio

The likelihood ratio of ARACNE network is calculated based on **Equation 2-8.** The likelihood ratio of ARACNE network is calculated based on **Equation 2-8,**

$$LR_i = \frac{P(I_i|P)}{P(I_i|N)}$$

Equation 2-8

where $P(I_i|P)$ is the probability of finding a true positive interaction (P) I in ARACNE network i and $P(I_i|N)$ is the probability of finding a false positive interaction (N) I in ARACNE network i. Computing likelihood ratios requires large datasets of both positive and negative examples (i.e. interactions that are respectively known to exist and not to exist). These are called Gold Standard Positive and Negative sets (GSP and GSN respectively). To generate the GSP, we extracted human interactions from the Transfac® Professional (TRANSFAC) (Matys, Fricke et al. 2003), BIND (Bader, Betel et al. 2003) and Myc (MycDB) databases (Zeller, Jegga et al. 2003). For defining a GSN, 500,000 gene pairs composed of a TF and a target were randomly generated, excluding pairs where the two genes are involved in a GSP interaction. GSP interactions are then restricted to interactions showing statistically significant mutual information in the cell type specific gene expression profile (i.e. B cell or TCGA brain data used in our research) used for generating ARACNE networks.

2.2.7 ARACNE Regulon Enrichment Analysis with Odds Ratio

In the gene knock-down experiment, differentially expressed genes were first identified in TF knock-down experiments (FDR < 0.05). ARACNE regulon (a.k.a. TF targets) was then compared to the differential expressed genes by an odds ratio (OR), which was calculated based on **Equation 2-9,**

$$OR = \frac{TP * TN}{FP * FN}$$

Equation 2-9

where TP is the number of overlapped genes between ARACNE regulon and differentially expressed genes, FP is the number of genes in ARACNE regulon but not found in differentially expressed genes, FN is the number of genes differentially expressed but not found in ARACNE regulon and TN is the number genes that are neither differentially expressed nor found in ARACNE regulon. The significance of OR was estimated by Fisher's exact test.

2.2.8 Comparing predicted TF regulons to ChEA database

ChIP Enrichment Analysis (ChEA) was a web-based interactive application for analyzing transcription factor targets from ChIP-based experimental data (Lachmann, Xu et al.). The backbone of this application is a database that contains the predicted TF targets in human and mouse genomes. All the TF-target interactions were either collected from the published literatures or predicted from raw ChIP-based data by using an in-house algorithm (For details, please refer to Lachmann *et al*'s paper). Currently, thirty-five human TFs with annotated targets are available from the ChEA database, which enabled taking advantage of the existing information and using them as a reference to compare against the predicted regulons from standard ARACNe and iARACNe.

2.3 Results

2.3.1 Transcription factors' mRNA expression level was not a good proxy to its protein activity when experiencing post-transcriptional regulation.

In the past decade, microarray expression data has been widely used in different aspects of research, including target discovery, biomarker determination, target selection, disease-subclass determination, *etc* (Butte, Tamayo et al. 2000). Most algorithms used microarray expression data as the input with the assumption that mRNA expression level of a gene could be used as a proxy to its protein activity. Although this assumption worked for most genes, it is not a reliable for genes that are post-transcriptionally regulated. In fact, post-transcriptional regulations were widely observed in eukaryotes, including splicing control, mRNA transcript stability, localization and translation (Day and Tuite 1998). Recent studies showed that microRNAs (miRs) (Filipowicz, Bhattacharyya et al. 2008) and RNA-binding proteins (RBPs) (Blencowe, Brenner et al. 2009) are both key players for gene post-transcriptional regulation. The average size of microRNAs is about 21 nucleotides. It was believed that microRNAs helped in regulating mRNA expression as well as in stabilizing mRNA segments in the cytoplasm. Another player, RBP, was found to be associated with hundreds of mRNAs during post-transcriptional regulation. We hypothesized that if transcription factors (TFs) experienced post-transcriptional regulation (**Figure 2-3**), the dissociation between their mRNA expression levels and protein activities could be observed.

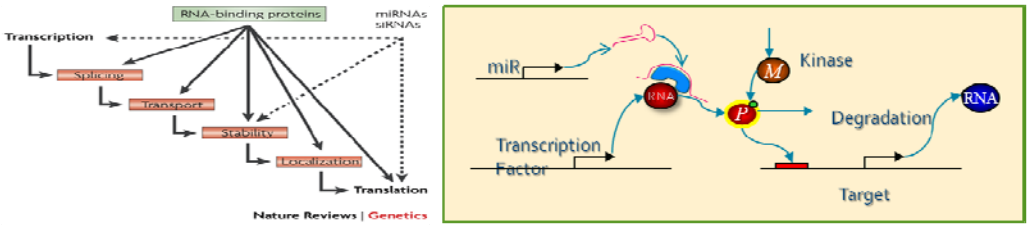


Figure 2-3 Demonstration of gene post-transcriptional regulation

We selected transcription factor CEBPb to test this hypothesis. CEBPb has been shown to play a very important role in breast cancer development (Zahnow 2002) and was recently identified as one the master regulators for breast cancer biogenesis (Lim, Lyashenko et al. 2009). Based on previously laboratory research, a set of genes that either positively or negatively modulated the interactions between CEBPb and its targets were identified using MINDY algorithm (Wang, Saito et al. 2009) (**Table 2-1**). These genes were referred as modulators in the paper.

Table 2-1. Potential CEBPb modulators

	ZWINT	HMBG2	RFC4	CCNA2	ECT2	TYMS	GTSE1	RANBP1	CDC42	PRDX4	PLCL2	RASL12	CREBL2
CEBPb	+	+	+	+	+	+	+	+	+	+	+	+	+
	EZH2	AURKA	DDR1	GPSM2	CNIH4	NPAS2	CD83						
	-	-	-	-	-	-	-						

If mRNA expression level was a good proxy to protein activity, we should observe a strong correlation between CEBPb's mRNA expression and its protein activity, with or without its modulator. In other words, if there was no change at the mRNA expression level, the protein activity should remain the same. TYMS and GTSE1 are two

modulators of CEBPb that positively regulated the interactions between CEBPb and its target genes in SNB-19, a human glioblastoma cell line. In the same cellular context, CD83 negatively regulated the interactions between CEBPb and its targets. By silencing TYMS, GTSE1 and CD83, respectively, the mRNA expressions of these three modulators were almost completely eliminated after 48hrs (**Figure 2-4 A**). When either of them was present, although CEBPb's mRNA expression level remained unchanged, their protein activities were either greatly reduced or increased (**Figure 2-4 B, C**). When TYMS was silenced, the protein activity of CEBPb decreased by 60%. When GTSE1 was silenced, CEBPb protein activity decreased to 50%. And when CD83 was silenced, CEBPb protein activity increased about 30%. These results strongly supported our hypothesis and suggested that mRNA expression level was not a reliable estimation of the transcription factor's protein activity, when post-transcriptional regulation was involved.

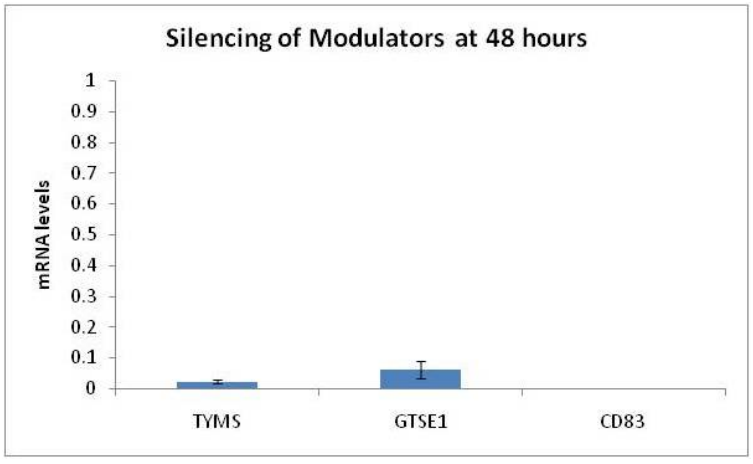


Figure 2-4 A. Silencing CEBPb modulators, TYMS, GTSE1 and CD83, in SNB-19 cells

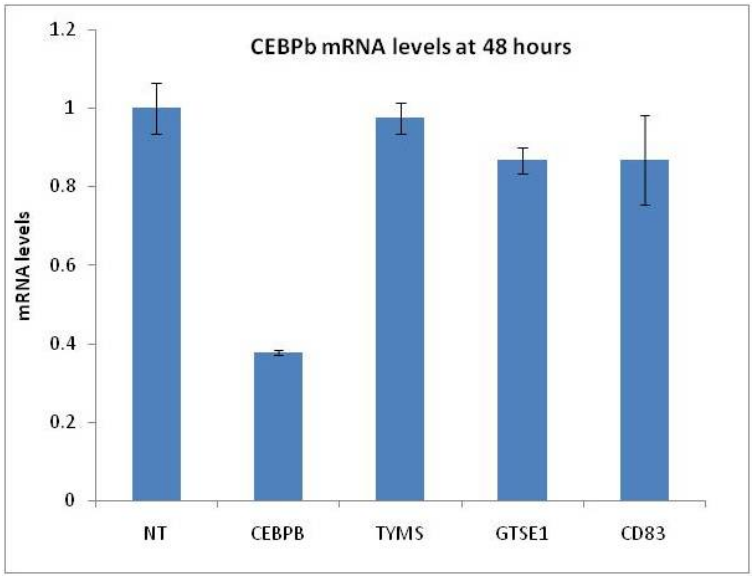


Figure 2-4 B. mRNA of CEBPb remained unchanged after silencing modulators. Y-axis represents the mRNA level of CEBPb after silencing corresponding modulators.

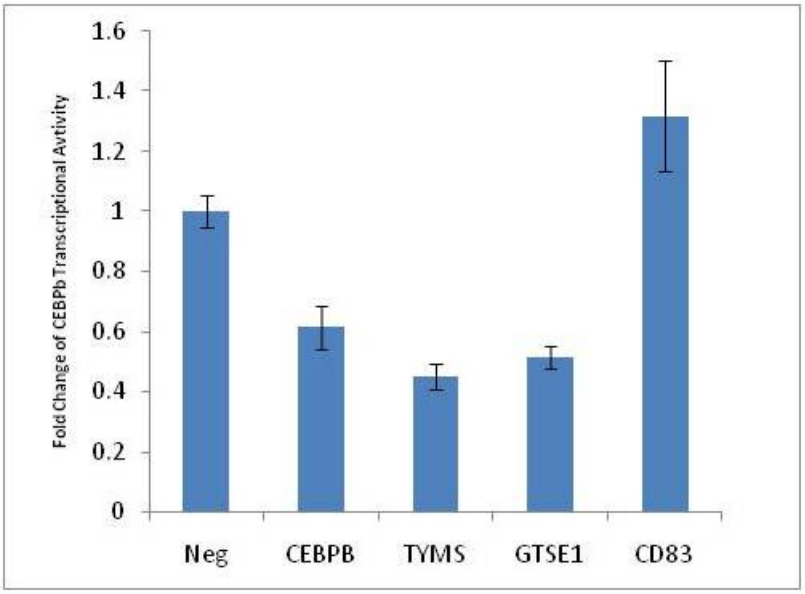


Figure 2-4 C. Silencing CEBPb modulators affected CEBPb proteins activities.

Since mRNA expression was not a good proxy for TFs that experienced post-transcriptional regulations, we hypothesized that if a more direct proxy to TF protein

activity could be used instead of mRNA expression, the predicted results would be more reliable. Previous work by Gao *et al* (Gao, Foat et al. 2004) and Youn *et al* (Youn, Reiss et al.) have shown that TF protein activity could be quantitatively modeled by their target genes' mRNA expressions. But the limitation of their approaches is that binding information of TF and target is required. Although relatively abundant in lower species, such as yeast genome, this binding information is hard to collect in higher species, such as human genome. Our algorithm addressed this question from a different angle. If all TFs' protein activities could be experimentally measured in large scale, we would use that information to reconstruct transcriptional networks. Unfortunately, genome-wide TF protein activity measurement method is not available. The following is a brief overview of the whole process. A TF gene is first transcribed into mRNA and mRNA is then translated into protein, which regulates the expression of all its targets (**Figure 2-5**). The more dynamic the regulated target genes' expressions are, the more active the TF protein. Thus, the overall mRNA expression performance of target genes is a more direct indicator of TF's protein activity.

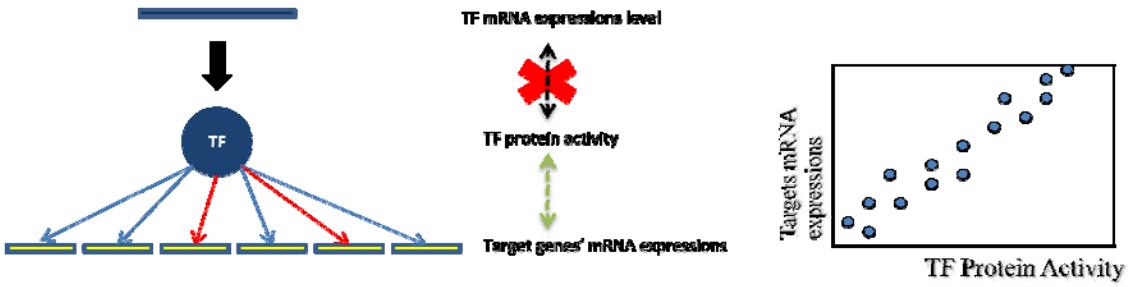


Figure 2-5. TF target genes' overall mRNA expression directly reflect TF protein activity

Therefore, the following hypotheses were made:

1) TFs protein activities could be estimated from the overall performance of their targets' mRNA expressions.
2) Transcriptional regulatory network could be reconstructed by using these newly estimated TF protein activities instead of TF mRNA expressions
3) The newly constructed network should be more reliable and robust than the one constructed simply from TF mRNA expressions.

To estimate TFs protein activity, the GSEA approach was applied. In the original GSEA setting, if the set of interested genes was from one biological condition (treatment) versus another (control), the interested genes are expected to be enriched in the gene set that were highly differentially expressed in the treatment condition compared to the control condition. This idea was transformed in the experiment design. A transcription factor's regulon is treated as the interested gene set and the enrichment score calculated from the regulon is considered as the activity of the TF. The more accurate the predicted regulon is, the more reliable the estimated TF protein activity will be. This idea was implemented by iteratively calculating TF protein activities and replacing original TF mRNA expressions in the network construction. Compared to the standard ARACNe which only uses TF mRNA expressions, the new approach is called iARACNe (please check **Materials and Methods** section for details). The diagram of iARACNe is demonstrated in **Figure 2-6**.

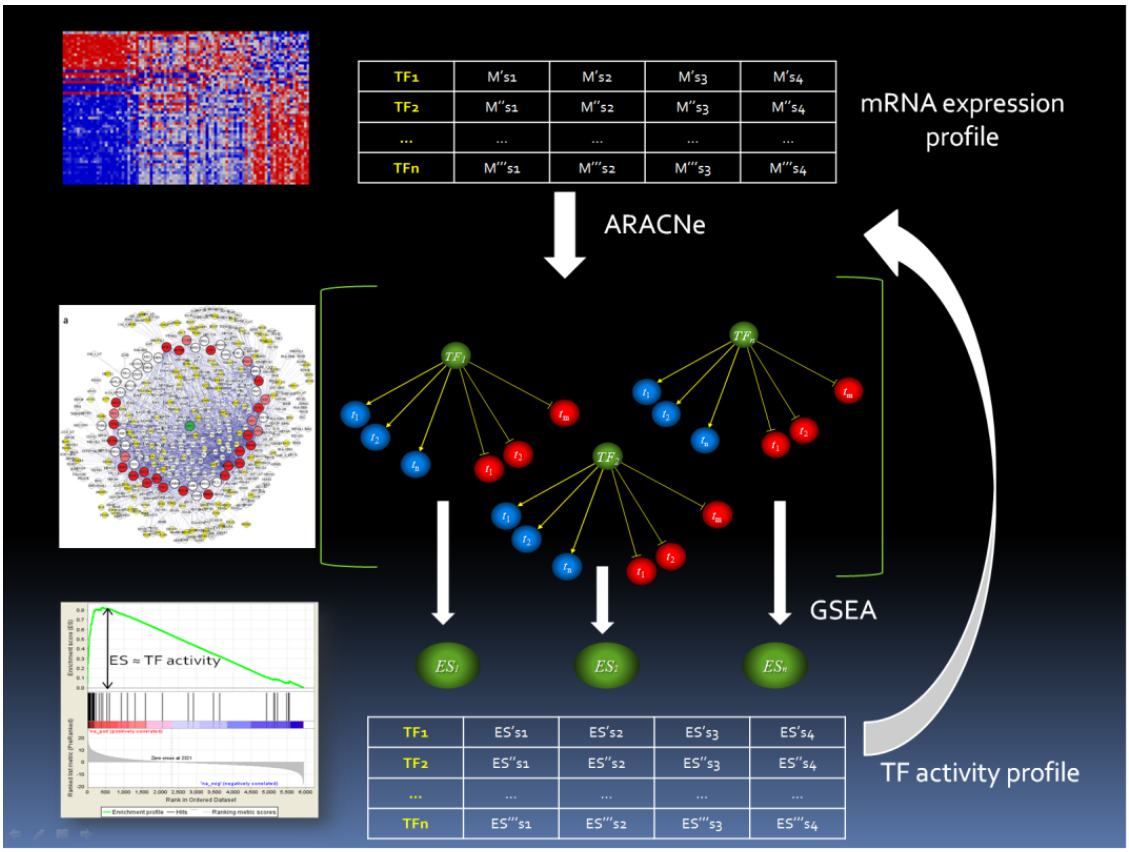


Figure 2-6. iARACNe diagram

2.3.2 Reverse engineering transcriptional networks with standard ARACNe and iARACNe

Glioblastoma is the most common and aggressive brain tumor. More than 50% of primary brain tumor cases belong to this category. Although the current occurrence rate for glioblastoma is not very high, about 2 to 3 cases per 100,000 people in western countries, the survival rate is very low. On average, patients diagnosed with glioblastoma die within three months if not treated, or they could survive 12 months if treated. Currently known GBM biomarkers didn't well separate high-grade gliomas patients from

low-grade gliomas. Therefore, there is a strong need to identify new biomarkers that enable a better prognosis before the treatment. In this study, public available gliomblastoma gene expression data from TCGA which contained 338 patients and reconstructed a genome-wide network for GBM transcriptional interactions was used. This network would serve as the basis for researchers to identify master regulators that control the transformation of GBM. In addition to GBM data, B-cell leukemia data was worked on. There were 201 samples from nine B-cell phenotypes including diffused large B cell lymphoma (DLBCL), follicular lymphoma (FL), chronic lymphocytic leukemia (CLL), memory B-cell, naïve B-cell, and so on. For more detailed information on the phenotype data please refer to Lefevre *et al* (Lefebvre, Rajbhandari et al.). Two cancer data sets were used to constructed transcriptional regulatory networks using standard ARACNe and iARACNe, respectively.

To reverse engineer regulatory networks, we first applied z-score transformation on mRNA expression data. In brief, gene expression intensity score under each sample condition was normalized by the mean of the intensities and the standard deviation across all sample conditions of that particular gene (**Equation 2-5**) (please check **Materials and Methods** section for details). Z-score transformation enabled us to place gene expressions from different samples into the same scale for comparison. In another word, each sample was considered as the treatment condition and the mean of all samples was considered as the control condition. Because of this normalization, when GSEA was applied later, all calculated enrichment scores were comparable to each other.

The basic properties of the constructed networks were compared, such as the number of interactions (edges), the distribution of the edges for TFs, and the distribution of hubs. In the B-cell data, there were 15252 probe ids, in which 1389 were transcription factors representing 802 unique TFs. The regulatory network constructed using standard ARACNe (R1) contained 244078 probe ids based TF-target interactions, representing 193,830 unique entrez id interactions. 1765 TF probe ids regulated at least 20 targets, either positively or negatively. The regulons of these TFs were used to calculate the protein activity of TFs'. Calculated TF protein activities were then applied to iARACNe. The first regulatory network constructed by iARACNe (R2) contained 56192 probe ids based TF-target interactions, representing 46403 unique entrez id based interactions, about one-fourth of the size of R1. 1129 TF probe ids regulated at least 20 targets. We then built the network based on the interactions from R2 network, named R3. In R3 network, total 57908 TF-target interactions were identified using probe ids, while when entrez ids of TF and target were used, the number of interactions dropped to 49296. 1005 TFs probe ids regulated at least 20 targets. In GBM data, we did the same analysis. We showed that there were 166414 interactions in R1 network, 42775 interactions in R2 network and 44609 in R3 network. 1489 TF probe ids in R1 network regulated at least 20 targets, while only 902 and 630 TFs met this criterion in R2 and R3 networks, respectively. As for the hubs, although most TF hubs remained the same, the number of edges they have changes. In B-cell data set, the largest TF hub in R1 network contained more than 900 edges (a.k.a. predicted targets). While in R2, and R3 networks, the largest hub only had about 200 edges. The same pattern was observed in GBM data as well. The

biggest hub in R1 networks had close to 500 interactions, but less than 150 in R2 and R3 networks (**Table 2-2**). The edge distribution in each network was computed (**Figure 2-7**).

Table 2-2. R1, R3 and R3 network features comparison

Data Set	Network	Total Edges	Largest Hub
B-cell	R1	193,830	955
	R2	56,192	132
	R3	57,908	228
GBM	R1	166,414	483
	R2	42,775	129
	R3	44,609	130

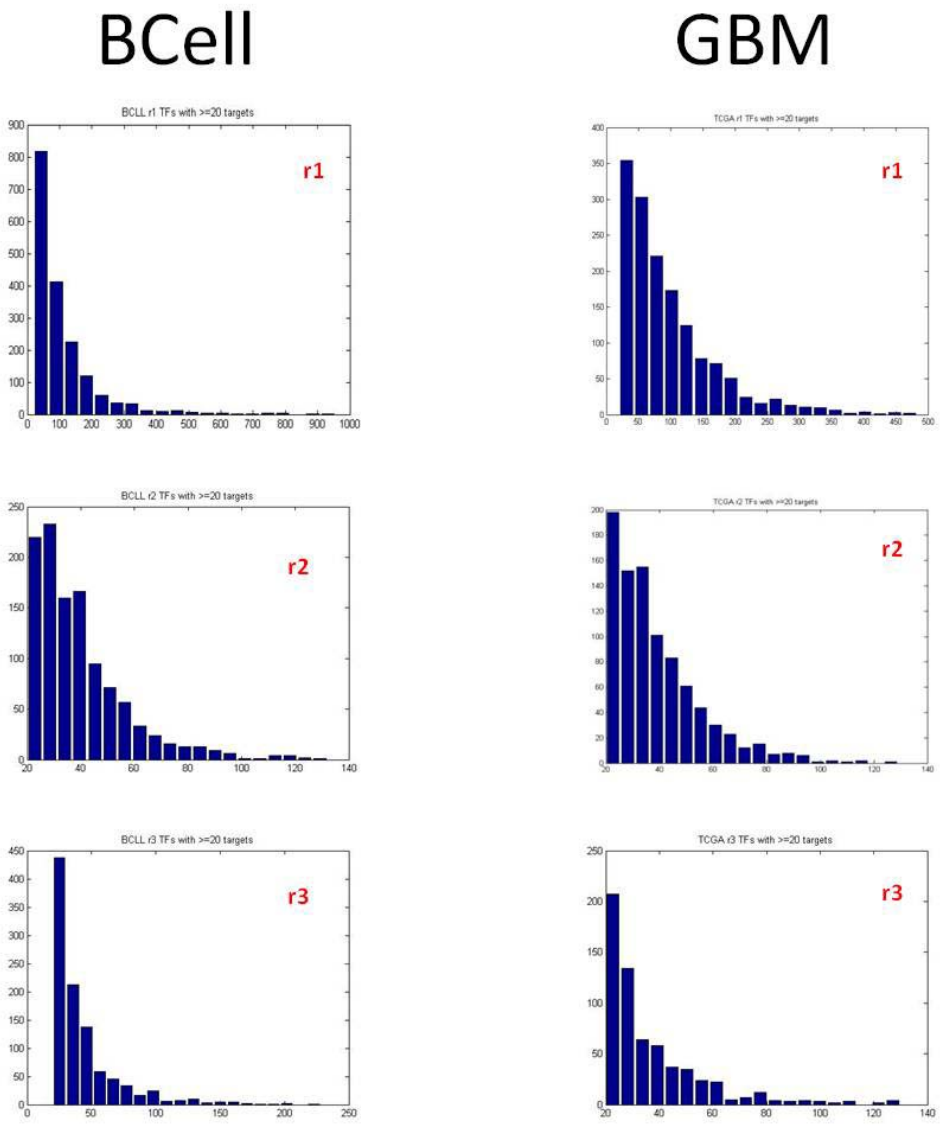


Figure 2-7. TF targets distributions in R1, R2 and R3 networks

The TFs in inferred networks were also compared based on their overall performances across all samples. We calculated Spearman correlations between TFs in the R1 network and TFs in the R2 network (R1 *vs.* R2), R2 *vs.* R3 and R1 *vs.* R3. In both B-cell data and GBM data, we showed that TFs' profiles in the R2 and R3 networks were very similar to each other, while different from those in the R1 network (**Figure 2-8**).

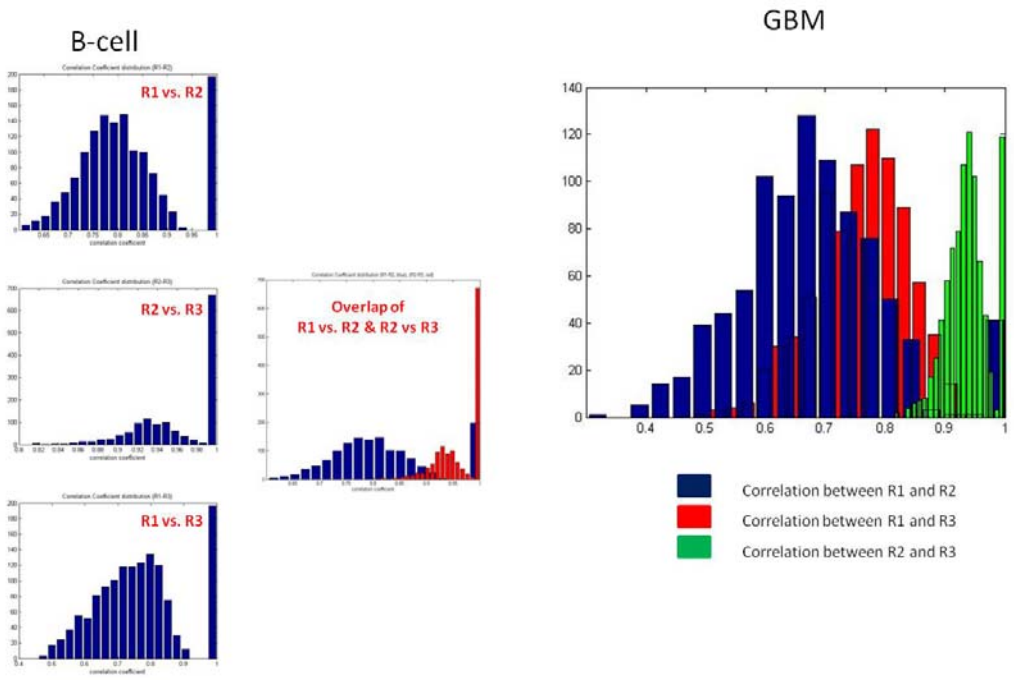


Figure 2-8. Comparing TFs profiles in the R1, R2 and R3 networks

A more strict comparison of the interactions inside each network showed that 66% of the interactions in the R2 network were also found in the R3 network (as well as verse vice), while only 16% of interactions in the R1 network were found in the R2 network and 15% in the R3 network. Now, the new question was why there was a decrease in the total interaction numbers. And were the new networks generated by iARACNe better than the one inferred by standard ARACNe? To answer these questions, we evaluated our prediction both functionally and physically.

2.3.3 Comparing TFs' regulons from iARACNe and standard ARACNe networks to differentially expressed signature genes

In the above section, it was showed that compared to the standard ARACNe constructed network (R1), iARACNe constructed networks (R2, R3) had fewer interactions. To test whether this decrease of the number of interactions improved the accuracy of the network prediction, several gene knock-down experiments were performed in the lab. The regulons from the R2 and R3 networks were expected to be more enriched with differentially expressed genes (a.k.a. signature genes) than those from the R1 network. STAT3 was selected to test our hypothesis. STAT3 plays an important role in different cancers and was recently identified as one the master regulators of breast cancer (Lim, Lyashenko et al. 2009) and human brain tumor (Carro, Lim et al.). STAT3 knock down experiments were performed in both human brain tumor initiating cells (BTIC) and SNB19 human GBM cell line by lentiviral-mediated shRNA transduction. Total 832 genes were differentially expressed (FDR $\leq$ 0.05) between 11 non-target shRNA control transduction samples and 11 STAT3 shRNA transduced in BTIC. iARACNe regulons showed a stronger overlap with the STAT3 knock-down differentially expressed genes than the standard ARACNe regulon. Moreover, the increase in the Odds Ratio cannot be achieved simply by selecting a subset of targets from the ARACNe regulon showing the highest mutual information. The improvement of each iteration *i* versus previous regulons *i-1* was estimated by comparing the Odds Ratio for regulon *i* with a null distribution generated by selecting a subset of target genes of same size as regulon *i* from the previous regulon *i-1* at random 10,000 times. It was shown that Odds Ratio from the R2 and R3 networks almost doubled that from the R1 network (**Table 2-3**). Similar increases were observed in the STAT3 knock-down experiments in SNP19 human glioblastoma cell line. In addition to STAT3, the regulons

from R1, R2 and R3 networks were also evaluated to previously predicted BCL6 targets (Basso, Saito et al. ; Polo, Juszczynski et al. 2007). In Basso *et al*'s paper, they identified 1207 BCL6 target genes in the germinal center (GC) B-cells by comparing to naïve and memory B-cells and ChIP-on-Chip experiment results. The comparison between the regulons from the R1, R2 and R3 networks showed that regulons from R2 and R3 had much better overlaps with identified BCL6 targets (**Table 2-3**). Another BCL6 target set was from Polo *et al*'s 180 BCL6 target genes in DLBCL. Again, R2 and R3's regulons have an improved overlap with this target set comparing the R1 regulon (**Table 2-3**). Overall, from both in house STAT3 knock-down experiments in brain tumor cell line and the external BCL6 targets in B-cell, the same improvements in iARACNe inferred regulons were observed, compared to the standard ARACNe regulons.

Table 2-3. Regulon enrichments analysis for R1, R2 and R3 networks

TF	Cell Line	Network	Regulon Size	OR	OR p-value	top OR	top OR p-value	ΔOR	ΔOR p-value
STAT3	BTIC	R1	572	2.562	7.96E-08				
		R2	83	4.369	1.17E-04	1.435	0.314	1.705	0.0242
		R3	71	3.619	0.00283	1.693	0.224	1.413	0.1023
	SNB19	R1	572	2.972	1.99E-11				
		R2	83	3.968	0.00026	3.151	0.00381	1.335	0.0999
		R3	71	2.829	0.0167	2.385	0.0495	0.952	0.416
BCL6	GC	R1	143	3.229	6.29E-07				
		R2	74	6.093	3.30E-10	2.171	0.0214	1.887	<10-4
		R3	71	5.307	2.70E-08	2.28	0.016	1.643	0.0011
	DLBCL	R1	143	6.128	3.98E-07				
		R2	74	8.664	1.00E-06	3.897	0.0122	1.414	0.03
		R3	71	10.235	6.69E-08	4.075	0.0103	1.67	0.0044

2.3.4 Validating iARACNe and standard ARACNe networks with using TRANSFAC and BIND interactions.

We have shown that iARACNe predicted regulons had a better enrichment with the genes that were functionally regulated by the given TF. Now, it is interested in understanding the overall performance of the networks. To evaluate the accuracy in these networks, the interactions from TRANSAC (Matys, Fricke et al. 2003) and BIND database were chosen as the gold standard. These interactions were compared to the array platform that was used for the experiments and the interactions which were also found in the array platform were selected. Based on this criterion 455 interactions were selected for B-cell data and 319 for GBM data set. These selected interactions were considered as gold standards. For negative controls, since there was no real negative pair information available, 500000 random pairs were selected. To evaluate the accuracy of the networks, we applied Network Likelihood Ratio (NLR) approach. In brief, our predictions were compared to the gold standard to identify true positive interactions and were compared to negative controls to identify false positives. NLR was defined as the ratio of identifying true interactions versus identifying false interactions (please check **Materials and Methods** for more details). The comparison showed that NLPs from R2 and R3 networks were almost two-fold of NLP from R1 in both b-cell and GBM data sets. Further analysis showed that this increase was due to the dramatic reduction of false positive predictions in the network (**Table 2-4)**. Although the percentage of true positives decreased about three-folds in iARACNe networks, the percentage of false positives decreased about six-folds. This suggested that the biggest advantage of iARACNe algorithm over standard ARACNe might be due to the fact that large numbers of false predictions were eliminated when TF protein activities were used during network construction.

Table 2-4. Comparison to TRANSFAC_BIND interactions

Data set	Network	Gold Standard	# of TP	# of FP	NLR
B-cell	R1	445	61	3176	22
	R2		23	517	43
	R3		20	638	37
GBM	R1	319	66	12996	8
	R2		21	2837	12
	R3		14	2250	10

2.3.5 Comparing iARACNe and standard ARACNe regulons to targets predicted from ChEA

All the interactions predicted in iARACNe and standard ARACNe were functional relationships between TFs and target genes, based on the correlations of their mRNA expressions. On the other hand, ChIP-based approaches detect the physical interactions between TFs and target genes by locating the TF binding sites on the target genes. Ideally, the true interactions should be the ones identified by both functional prediction and physical prediction. ChEA is the latest ChIP-based TF-target database developed by Ma'ayan's group (Lachmann, Xu et al.). In brief, they collected all publicly available ChIP-related experimental data, such as ChIP-on-Chip, ChIP-seq, ChIP-PET and DamID. For a given TF, if its ChIP targets were available from the published literatures, they imported them into the ChEA database. Otherwise, TF targets were predicted from raw ChIP-X data using their own algorithm. Currently, 57 human and 78 mouse TFs were stored in the ChEA database. This database allows an opportunity to

connect functional relationships to physical interactions. Therefore a comparison of the predicted regulons from regulatory network was made to ChEA human TFs targets obtained from ChIP-X experiments. Twenty-two human TFs in ChEA database were also found in B-Cell R1 and R2 regulatory networks, and 21 human TFs in R3 network. 21342 TF-target interactions associated with these 22 TFs or 15960 TF-target interactions associated with 21 TFs were defined in ChEA. In the GBM data set, 22 human TFs were found in all three inferred networks and 28282 TF-target interactions were found in the ChEA database. It was shown that, consistent with previous comparisons with TRANSFAC-BIND interactions, networks generated by iARACNe always had a better overlapping with ChIP-X based interactions (**Table 2-5**). The NLRs in R2 and R3 networks doubled that in R1 network in both datasets.

Table 2-5. Network Likelihood Ratios of R1, R2 and R3 comparing to ChEA database

Data set	Network	Gold Standard	# of TP	# of FP	NLR
B-cell	R1	21342	336	3176	24
	R2	21342	107	517	48
	R3	15960	98	638	48
GBM	R1	28282	386	12996	5
	R2		167	2837	10
	R3		164	2250	12

2.4 Discussion

One of the major goals of systems biology is to understand the regulation of gene expression. Although gene expression is controlled at different stages, control of mRNA transcription by transcription factors (TFs) has so far captured the greatest attention. The goal of this research was to further improve our understanding of transcriptional regulation mechanisms. I was especially interested in tumorigenesis and in distinguishing cancer sub-phenotypes based on their transcriptional regulation programs. To address these questions, we must identify the transcription factor genes that regulate these programs and are thus responsible for tumorigenesis and the emergence of distinct tumor subtypes. These genes are termed master regulators.

Can this question be addressed computationally, using a regulatory model? Clearly, algorithms designed to address this problem will greatly benefit from the ability to assemble reliable transcriptional regulatory network models so that each TF can be studied based on the targets it regulates. This research is targeted to and therefore focuses on improving the accuracy of transcriptional regulation network and the mechanistic understanding of TF-target regulation. Currently, most network-inferring algorithms take microarray expression data as their inputs to reconstruct the network. Their assumption is that gene mRNA expression level can be used as a proxy to its protein activity. This assumption has always been questioned by biologists, because it is frequently violated in the cell. Post-translational (e.g. by kinases, phosphatases, ubiquitin conjugating ligases), post-transcriptional (e.g., by microRNAs and mRNA binding proteins), and translational

regulation (e.g. by ribosome binding proteins) break the direct link between the mRNA expression of a TF and its activity at the protein level. The current study has shown, for instance, that post-translational regulation of the CEBPb TF significantly affects its protein activity even though its mRNA does not change (please check section 2.2.1 for details).

Ideally, we would like to experimentally measure the concentrations of all active TFs' under different conditions. But due to low expression level and fast post-transcriptional modification, it is very difficult to directly measure active TFs' concentrations. Since currently there is no method that enables us to measure genome-wide TF protein activity, if mRNA should not be used as a direct proxy to protein activity, what should be used? To obviate this problem, the current research developed an approach that iteratively infers TF activity levels by using the predicted TF regulons (aka target genes). This approach was named iARACNe and based on two features. First, it is an extension of the well established ARACNe algorithm, which means that it took advantages of the ARACNe constructed network. Second, the active TFs' activities were quantitatively and iteratively estimated based on revised TF regulons, thus providing more accurate estimates than the mRNA.

The idea of using target genes' observed expression data to infer TFs' activities (TFAs) has gained significant attention in recent years. Several approaches have been proposed and implemented. Based on the strategies from which TF activity was inferred,

these approaches can be broadly placed into two categories. Approaches in the first one are mainly based on a modified linear regression model(Liao, Boscolo et al. 2003; Gao, Foat et al. 2004; Boulesteix and Strimmer 2005). *A priori* knowledge of network connectivity (network topology), was required in these approaches. Topology was generally inferred either from ChIP experiments or from DNA binding motif analysis. The goal of these methods was the assembly of a transcription factor activities matrix, which represents the activity of each TFs in each sample for which expression data is available. Such a strategy has limited applications in human genomes, due to the lack of ChIP data for most TFs, limited knowledge of binding motif information, size of regulatory regions, non-linear nature of TF-target relationships, and the cell-context-specific nature of transcriptional regulation in multi-cellular organisms.

iARACNe addresses these challenges by using an ARACNe-inferred network as an initial estimate of the TF targets and then by using this first inference to compute a better estimate of TF activity, based on target expression in each sample. Since ARACNe has been shown to infer reasonably accurate networks for mammalian cells, the initial ARACNe network connectivity constitutes a reasonable starting point to infer TF activity and reconstruct a more refined connectivity. A significant disadvantage of the linear regression model is that it significantly affected by the contribution from false-positive targets. The GSEA algorithm that iARACNe uses to infer TF activity, on the other hand, is much less affected by false-positives, which tend to be uniformly distributed and thus do not contribute to the normalized enrichment score.

A second category of TF activity inference approaches uses non-linear models, for instance based on Bayesian inference or differential equations (Asif, Rolfe et al. ;

Opper and Sanguinetti ; Barenco, Tomescu et al. 2006; Sabatti and James 2006; Sanguinetti, Lawrence et al. 2006). Although these models are biologically more plausible, they cannot learn regulatory models for large networks because of over-fitting and exponential growth in regulatory programs and are best used for small, well-defined sub-networks.

To evaluate the networks constructed by iARACNe and standard ARACNe, this research compared them to multiple gold standards. In STAT3 knock-down experiments, it was shown that regulons from iARACNe network had a better enrichment with differentially expressed genes compared to regulons from standard ARACNe. By comparing to the known interactions in TRANSFAC and BIND databases, It was demonstrated that the overall network reliability (a.k.a network likelihood ratio) in iARACNe networks almost doubled that of standard ARACNe networks. Compared to ChIP data based interactions, the same pattern was observed. Overall, it was demonstrated that by introducing TF protein activities into network reconstruction, we can achieve a more reliable network than the one constructed merely with mRNA expression.

There are, however, some relevant limitations of iARACNe too. First, because iARACNe used gene set enrichment analysis to calculate the TF activity from its target genes, at least 20 target genes were required to ensure the statistical reliability of the computed enrichment score. Therefore, some TFs that has fewer than 20 target genes

would not be updated to their protein activity, although iteration solved this problem most of the time. Second, the TF regulon predicted by iARACNe is much smaller than the one from standard ARACNe. Although the accuracy of the regulon increased, some known interactions were missed. Therefore, the best application for iARACNe is for researchers to identify the most reliable candidate for further investment. If the goal is to identify all possible interactions, standard ARACNe is a better choice.

Because iARACNe enabled an estimation of TF protein activities genome-wide from mRNA expression data, it opened a new gateway for research in this field. We can now associate TF activities to different tissues at different disease conditions.

Chapter 3

TF-centric motif discovery approach

3.1 Introduction

To understand transcriptional regulation, we first need to know the key player in this process, transcription factor (TF). By binding to the specific sites on the genome and collaborating with other enzymatic complexes, such as RNA polymerase, TFs precisely control the expressions of variety of genes at appropriate times and locations. When this sophisticated regulation is lost, we develop all kinds of diseases. For example, a lot of human development disorders are associated with dysfunctional TFs (Boyadjiev and Jabs 2000) and overrepresentation of TFs has been shown linked to oncogenesis (Furney, Higgins et al. 2006). For a good review about human diseases and transcriptional regulatory elements association, please refer to Maston *et al* (Maston, Evans et al. 2006). In the previous section, we have showed that a transcriptional regulatory network could be reverse engineered using genes mRNA expression profiles collected from different biological conditions. In addition, the accuracy of the network could be further improved when TFs' protein activities were iteratively introduced during the construction process. The inferred network contains the interactions between TFs and their corresponding targets (a.k.a. TF regulon). Now, we moved forward to understand how TF regulates or selects their target genes.

In order to regulate its target gene, TF first needs to bind to the promoter of the gene. Therefore, it is rationale to scan the promoter region of genes for the binding sites of the given TF. Initially, about 2000 to 3000 sequence-specific DNA-binding TFs were estimated in human genome (Lander, Linton et al. 2001; Venter, Adams et al. 2001). More recently, based on mapping InterPro (Hunter, Apweiler et al. 2009) DNA-binding domains, GO database predicted 1052 TFs genes for human (Ashburner, Ball et al. 2000). Although TFs are extremely important, of all these TFs, only 62 have been experimentally validated for their DNA-binding signatures and regulatory functions (Vaquerizas, Kummerfeld et al. 2009). TRANSFAC database (Matys, Fricke et al. 2003) has the largest collection of TF motifs, but less than 400 human TFs were annotated there. In addition, under different biological conditions and in different tissues, TFs might regulate different set of targets. Therefore, to better understand the mechanism of how TF regulates its targets in human, we need

1) A comprehensive binding motif profiles for all available human TFs
2) Associate TF binding motifs with different cellular contexts

Our study was an attempt toward this direction.

Motif discovery has always been one of the major challenges in biology. For more than a decade, researchers from different groups around the world have developed numerous binding site discovery algorithms (Stormo and Hartzell 1989; Lawrence, Altschul et al. 1993; Bailey and Elkan 1994; Bailey and Elkan 1995; Hughes, Estep et al. 2000; Sinha and Tompa 2003; Smith, Sumazin et al. 2005). Comparing to scan for known

motifs, *de novo* motif discovery is even hard, due to the large number of false positive prediction, especially for higher species genomes, such as human (Tompa, Li et al. 2005). Internally, the nature of TF binding motifs makes them hard to detect. As we know, TF binding motifs are small pieces of sequences, usually ranging from 5 to 12 base pairs. Looking for these small pieces in human genome which contains more than 3 billion base pairs (Birney, Stamatoyannopoulos et al. 2007) is literately like looking for a needle in not one, but thousands of haystacks. These binding sites are also highly degenerate. Not all the nucleotides in the motif are conserved, but rather very limited number of them are highly conserved and play the key role (Stormo 2000). This feature enables TF to have multiple possible binding sites for different target genes. Good for the cell, but bad for scientist. In addition, nucleotides in the motif are not completely independent of each other (Man and Stormo 2001). Although most motif algorithms disregard this fact, it does affect the prediction in some way. Finally, some TFs functions are not conserved, it has also been shown that TFs do not always regulate their target genes in all tissues or in different species (Vaquerizas, Kummerfeld et al. 2009). All these internal factors make *de novo* TF motif discovery different and generated a lot of false positive predictions. Externally, the major challenge for motif discovery is to obtain the reliable sequence set from which we can predict the given TF's binding site (TFBS). Let's first brief explain different ways of defining TFBS. TF binding motif can either be represented as a consensus, such as words (Sadler, Waterman et al. 1983; Blanchette and Sinha 2001) or regular expressions (Brazma, Jonassen et al. 1996; Califano 2000), or a position-weight matrix (PWM) (Benos, Bulyk et al. 2002; Bulyk, Johnson et al. 2002) or a position-specific scoring matrix (PSSM) (Stormo and Hartzell 1989; Lawrence, Altschul et al.

1993; Bailey and Elkan 1994; Smith, Sumazin et al. 2005). Each definition has its advantages and disadvantages. In our study, we selected PWMs to summarize TFBSs because validated PWMs are available from several sources (Matys, Fricke et al. 2003; Sandelin, Alkema et al. 2004), and they are suitable for *de novo* discovery as they provide a good tradeoff between binding site prediction accuracy and the required volume of training data needed (Smith, Sumazin et al. 2007). We study a variation on the original formulation of the motif discovery problem, which was introduced by Yoseph et al. (Yoseph, Gill et al. 2001). They discovered motifs that are enriched in a foreground sequence set against a control set, and the advantage of their approach was demonstrated using both regular-expression motifs and PWMs (Smith, Sumazin et al. 2005; Sumazin, Chen et al. 2005).

As mentioned above, there is a big gap between available human TF binding motif profiles and the number of known TFs. Only about 15% of TFs contain well characterized motif profiles. In addition, many motif discovery algorithms only worked well in lower species, such as *E.coli* and yeast, of which the genomes are relatively simple, small and with low gene regulation complexity. When applied to higher species, such as mouse and human genomes, the recalls of these algorithms were disappointing, only about 15% (Tompa, Li et al. 2005). This was due to the fact that gene regulations in higher species' genome are far more complicated than those in the lower species, in terms of number of TFs involved, the distance of TF bind sites to the TSS, the collaborations between different TFs in regulation one gene's expression and *etc*. Therefore, there is a need to improve the motif discovery accuracy for human genome.

Expression, binding, and cross-species conservation data have all been used to guide motif discovery methods. Co-expression with TFs was used to identify putative promoters that may contain binding sites for TFs and could then be analyzed for TFBS enrichment (Aach, Rindone et al. 2000; Conlon, Liu et al. 2003; Beer and Tavazoie 2004). Cross-species conservation was used to identify genomic regions that are more likely to be functionally important and thus enriched with TFBSs and other regulatory elements (Blanchette and Tompa 2002; Moses, Chiang et al. 2004). Finally, some of the most successful motif and TFBS discovery approaches use binding data and especially high-throughput chromatin immunoprecipitation (ChIP-chip and ChIP-seq) data to identify relatively short target DNA regions with high likelihood for binding-site presence (Smith, Sumazin et al. 2005; Kim, Abdullaev et al. 2007; Ward and Bussemaker 2008). However, due to limited antibody availability, cell-context specificity of transcriptional interaction patterns, and the associated cost, the assembly of complete binding site repertoires for the majority of TFs is not a viable option.

In our research, we show that a significant improvement in TFBS discovery can be achieved by using an integrative work-flow approach we call OmniMiner. First, we use ARACNe, a proven reverse-engineering algorithm (Basso, Margolin et al. 2005; Margolin, Nemenman et al. 2006; Margolin, Wang et al. 2006; Carro, Lim et al. 2009), to identify higher likelihood transcriptional targets, and we demonstrate that the inferred targets are more reliable than those predicted by co-expression. Our results suggest that by using ARACNe-predicted targets we significantly improve accuracy when compared

to the co-expression approach by removing false positives among high-confidence and especially among low-confidence co-expressed targets. Then, we identify cross-species conserved regions by combining linear-alignment and pattern-discovery (*alignment-free*) based approaches. Genome-alignment-based conservation (Siepel, Bejerano et al. 2005) can guide motif discovery (Xie, Mikkelsen et al. 2007) and help identify motifs and sites for some regulators, but it may also obscure sites that are not conserved linearly as is the case with binding-site turnover. We correct for this and show that combining the two approaches leads to significant prediction improvements. Finally, we use DME, a proven deterministic motif discovery algorithm (Smith, Sumazin et al. 2005; Kim, Abdullaev et al. 2007; Smith, Sumazin et al. 2007), to discover *de novo* TFBS motifs for specific TFs and their co-factors. In our experiments, the top OmniMiner *de novo* discovered motif matched a known motif for more than 15% of the TFs in our human B cell test set. OmniMiner's recall was over 30% when the criteria was expanded to include predictions where at least one of the top five motifs matched a known motif for the TF; we note that other top 5 significant motifs may describe the binding of a co-factor. In total, our results suggest that OmniMiner's performance on unaltered human promoters is better than the performance of methods described by Tompa et al. (Tompa, Li et al. 2005) on impregnated human promoters despite the fact that motif discovery in the former is widely considered to be more challenging.

To evaluate the performance improvement associated with better target selection and cross-species conservation, we assembled human promoter sets for genes predicted to be either co-expressed with or direct transcriptional targets of a representative collection

of TFs. To evaluate binding site enrichment, we measured the classification accuracy of verified TRANSFAC binding motifs associated with the TF (Matys, Fricke et al. 2003). We used binding site enrichment to compare recall rates across methods and to estimate the accuracy of *de novo* discovery methods. Then, we showed that while both our target-selection and cross-species-conservation methods improve our ability to discover bona-fide TFBSs for specific TFs, the greatest improvement arises from the integration of both methods. We compared our *de novo* motif discovery approach with GibbsModule (Xie, Cai et al. 2008), a method that was recently proposed as the state-of-the-art in integration of co-expression and cross-species conservation. While OmniMiner proceeds greedily, by identifying cross-species conserved regions in each promoter and patterns common to these conserved regions across promoters of inferred targets of a given TF, GibbsModule simultaneously identifies patterns conserved across species and across promoters of inferred targets. The simultaneous approach has the potential to maximize accuracy, but we show that OmniMiner's greedy approach produces significantly better results.

To support our estimate for prediction accuracy, we biochemically validated predictions for three TFs. Sites matching a known E2F1 motif were identified as the most enriched in predicted E2F1 targets and the second most enriched in JUND targets. Our validation confirms the presence of predicted E2F1 sites in promoters of predicted E2F1 targets, and it suggests that the majority of JUND targets are occupied by both TFs, which is consistent with the predicted co-factor role for E2F1. To demonstrate the accuracy of OmniMiner's *de novo* discovery, we validated predicted BCL6 binding sites in conserved regions of promoters of BCL6-predicted targets. Finally, to demonstrate

prediction accuracy using an external dataset, we tested *de novo* discovered motifs in promoters of predicted ZNF263 targets for enrichment in ZNF263-bound regions according to ChIP-seq (Farnham 2009). Our analysis showed that the three best *de novo* motifs are significantly enriched in ZNF263-bound regions.

3.2 Materials and Methods

3.2.1 ARACNe Network Inference

ARACNe is an information-theoretic method for identifying transcriptional interactions between TFs and their targets using gene expression profile (GEP) data. In brief, the algorithm first distinguishes candidate interactions between a TF and its targets by estimating the expression pairwise mutual information (MI). Interactions with significant MI values are retained (details can be found at ref 14). Then, ARACNe applies the Data Processing Inequality (DPI) theorem to eliminate the vast majority of interactions with significant MI values that are indirect and falsely predicted because of transcriptional interaction cascades. ARACNe with bootstraps uses bootstrap sampling during network reconstruction to non-parametrically assess statistical confidence for predicted transcriptional interactions. As a result, the built networks are more robust to both expression estimation and MI estimation errors. Dataset samples were randomly chosen with replacement and assembled into bootstrap datasets. In our experiments, 100 bootstrap datasets were generated and ARACNe was used to generate a set of bootstrap networks. Each bootstrap network contributed to a consensus network made of edges that were supported across a significant number of the bootstrap networks, where significance was measured using permutation testing with the null generated using shuffled networks and cutoff set to $p < 1e\text{-}7$ to correct for multiple testing.

3.2.2 Co-expression

We used 254 gene-expression profiles collected from a variety of homogeneous B cell phenotypes by Basso *et al* (Basso, Margolin et al. 2005) using the Affymetrix HG-U95A GeneChip® System; experimentally manipulated cell lines were excluded. TFs were selected among the genes represented on the HG-U95A microarray based on Gene Ontology annotation.

For each TF we identified (a) a set of co-expressed genes using Spearman correlation with a Bonferroni-corrected statistical threshold of 1e-4, and (b) a set of ARACNe inferred targets using a Mutual-Information-based Bonferroni-corrected significance threshold of 5e-2 and recorded positively correlated targets with each regulator. The Spearman correlation threshold was set low because higher threshold settings produced significantly larger target-gene sets and poor analysis results. ARACNe predicted TF target sets of size 30 or greater for seventy TFs with verified binding motifs in TRANSFAC. These TFs were used to compile our test set. The number of statistically co-expressed genes with each TF was significantly larger than the number of ARACNe-identified targets, and analysis results showed that ARACNe-identified targets are significantly more enriched with sites matching validated binding site motifs. To test if the disparity in site enrichment was related to the disparity in target set sizes, we selected the top *n* most statistically significant co-expressed targets for each

TF, where *n* was the size of the ARACNe target set for this TF. This test set makes the co-expression* set.

3.2.3 TF target sequences

We obtained 1500bp promoters for each target gene by selecting [-1000, 500] from Refseq transcription start site locations, eliminating intersecting promoters arbitrarily; we refer to these as the *conservation-free* sets. We masked repeats and coding exons to obtain the Masked Coding-exons and Repeats (*MCR*) set, which was used for computing pattern-discovery-based cross-species conservation. Alignment-based conservation was computed using 17-species PhastCons (Siepel, Bejerano et al. 2005). PhastCons requires a conservation probability parameter; we mapped the conservation probability parameter to DNA coverage proportion in order to achieve comparable statistics across regulators and conservation measures. A mapping of conservation probability to DNA coverage proportion for the seventy TFs is given in **Figure 3-1**. To add pattern-discovery-based conservation, we retrieved orthologous promoters for mouse, rat, chimpanzee, rhesus and dog. We used SPLASH (Hart, Royyuru et al. 2000; Sosinsky, Honig et al. 2007), a deterministic pattern discovery algorithm, to identify patterns across species after masking repeats and coding exons. When running SPLASH, we used eight-base windows for motif-seed discovery with a minimum six-base match within the window, and required a match across at least four species. SPLASH-identified conserved patterns were ranked by z-scores, and the top patterns were used to achieve a given DNA coverage proportion. Entire regions included in a sparse pattern were considered conserved. We combined alignment-based conservations at 10% DNA coverage with

pattern-discovery-based conservations at 10% coverage to construct *combined conservation target sequences (conservation*)*. Regions that were not considered conserved according to pattern-discovery-based or alignment-based conservation were masked out.

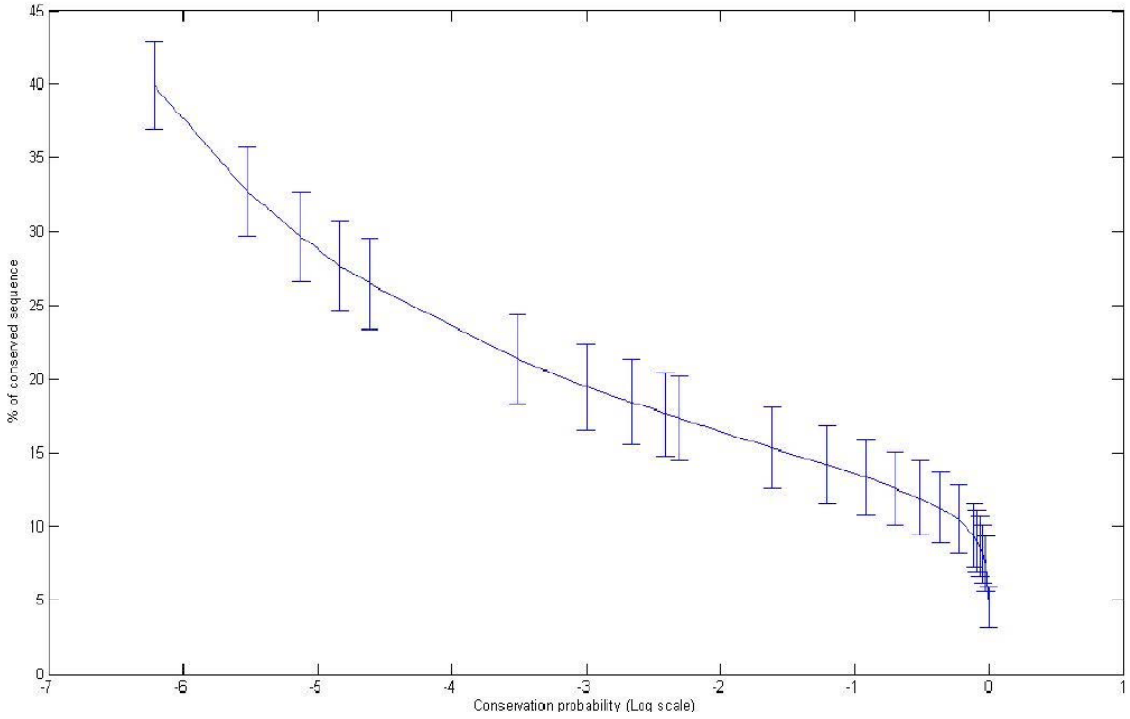


Figure 3-1. phastCons conservation probabilities and corresponding conservation-sequence proportions.

Our control set (background) was composed of 2000 non-overlapping promoters associated with randomly selected Refseq genes not identified as ARACNe or co-expression targets. These promoters were processed to obtain a background MCR set, alignment-free regions, and combined conservation sequences. When evaluating or discovering motifs enriched in a foreground set, we used the background set whose processing matched the processing of the foreground set.

3.2.4 Motif evaluation and discovery

De novo motif discovery was performed for TFs with significant binding site enrichment and for 20 TFs with no known binding characterization. We identified 103 TFs that activated at least 30 targets and had no known associated motifs. We ranked these TFs based on the number of PubMed abstracts containing the name of the TF (**Table 3-1**). *De novo* motif discovery was performed for the top 20 most cited TFs.

Table 3-1. Pubmed citations of TFs

TF	# of targets	pubmed hits
APC	33	12863
MSC	75	4848
HIF1A	46	3507
MLL	83	1615
RB1	39	1446
CEBPZ	71	914
MECP2	30	849
ID1	101	552
BCL6	52	515
VAV1	43	515
NFE2L2	36	459
FLI1	70	444
PAX7	84	382
EP300	31	378
NFATC1	34	314
MEF2C	71	283
NME2	107	222
PITX1	36	214
HOXD13	193	180
TBX1	33	152

Motif enrichment in foreground sets against background sets was measured using classification relative error rate (*err*), where relative error rate is computed as the average

of the false positive and false negative rates (Smith, Sumazin et al. 2007). Relative error rates were associated with p-values using permutation testing, where the indicator vector that assigns set membership to foreground or background is randomly permuted. When identifying discriminating motifs in a motif library, we assigned a p-value to an error rate by ranking it relative to the library's top error rates in 10,000 permutation tests. When assigning p-values to *de novo* identified motifs, we first generated 100 random foreground-background pair sets by permuting the indicator vector as described above. We then applied DME (Smith, Sumazin et al. 2005) to each of the 100 random foreground-background pair sets. In each permutation test, the score of the motif with the lowest relative error was recorded, and the resulting set of 100 relative error rates served as a null distribution against which we assessed the statistical significance of the *de novo* identified motifs from the original set. Motifs in the 95^{th} percentile ($p \leq 0.05$) are said to be statistically significant.

We used *matcompare* (Schones, Sumazin et al. 2005) with a similarity cutoff of 1.0 bit for motif comparison. DME (Smith, Sumazin et al. 2005) was used to discover enriched motifs of length 6, 8, and 10. Similar top motifs were merged using uniqmotifs (Smith, Sumazin et al. 2007). *GibbsModule* (Xie, Cai et al. 2008) was used to identify motifs of length 8, 10 and 12 with the default 300 iteration per execution.

3.2.5 Validation

We set out to validate ARACNe target predictions, TRNASFAC-based binding site predictions, co-factor binding predictions, and *de novo* motif discovery predictions. With consideration to anti-body availability, we chose to validate binding predictions for three TFs. The TRASNFAC E2F1 motif M00918 was identified as the most enriched motif in E2F1 targets, and the TRASNFAC E2F1 motif M00428 was identified as the most enriched motif in JUND targets. We validated BCL6 binding to sites identified using the top BCL6 motif candidate. Antibodies used for the study were anti-E2F1 (sc-251), anti-JUND (sc-74), anti-BCL6 (sc-585) and anti-GAPDH (sc-32233) from Santa Cruz Biotechnology.

Chromatin immunoprecipitation (ChIP) analysis was done in Ramos and MUTU-I cell lines by following the protocol described by (Kouskouti, Scheer et al. 2004). Ramos and MUTU-I cells were maintained in Iscove's modified Dulbecco's medium supplemented with 10% FBS and antibiotics. The soluble chromatin fraction was immunoprecipitated with anti-E2F1 or mouse IgG control antibody (MUTU-I), anti-JUND or rabbit IgG control antibody (MUTU-I), and anti-BCL6 or mouse IgG control antibody (Ramos). The immunoprecipitated DNA was reverse cross-linked and purified by phenol-chloroform. The chromatin fragments from two independent experiments were pooled and the amount of DNA immunoprecipitated by an individual antibody was assessed by real-time PCR in 7300 Real-time PCR System using Power SYBR Green (Applied Biosystems).

Two ZNF263ChIP-seq replicate experiments and IgG control in K562 cell line were obtained from UCSD ENCODE Data Release: Transcription Factor Binding Sites from Yale/UC-Davis/Harvard. MACS (Zhang, Liu et al. 2008) was used under default settings to predict (1) ZNF263 and (2) IgG bound regions. We used the top 500 ZNF263-bound regions as foreground, and as background we selected the top IgG bound regions to equal the total DNA of the foreground. Motif training and binding site detection followed the process described above, and enrichment p-values were calculated using Fisher exact test, comparing detection rates in each set, with Bonferroni correction for multiple testing.

3.3 Results

3.3.1 Use of reverse-engineering methods to identify TF targets

Co-expression has been widely used to infer regulatory interactions between TFs and their targets (Wasserman, Palumbo et al. 2000; Zhu, Pilpel et al. 2002; Wang and Stormo 2003; Beer and Tavazoie 2004), but co-expression alone is not sufficient for determining direct interactions. Gene sets that are co-expressed with a TF are generally enriched in its targets but also contain a large proportion of non-target and indirect targets, which substantially dilute enrichment. Regulatory networks reverse engineering algorithms like ARACNe, on the other hand, attempt to use additional properties of the data to identify genes that are more likely to be direct transcriptional targets of the TFs. Specifically, ARACNe uses the Data Processing Inequality theorem of mutual information, as well as direct knowledge of TF identity, to remove candidate regulatory interactions that are likely to be of an indirect nature (Margolin, Nemenman et al. 2006; Margolin, Wang et al. 2006). We used ARACNe with 100 rounds of bootstrapping to construct a regulatory network from 254 human B-cell gene-expression profiles (see ARACNe Network Inference in **Materials and Methods**). Since activation and repression can be mediated by distinct co-factors and binding sites (Phan, Saito et al. 2005), we concentrated strictly on targets predicted to be activated by the TF; these constitute the majority of the interactions in the reverse-engineered regulatory network and extension to repressed subsets is straightforward. As a representative TF set for performance analysis, we selected the 70 TFs with known DNA binding motifs in TRANSFAC (Matys, Fricke et al. 2003) that were predicted by ARACNe to be positive regulators of at least thirty targets, thus allowing appropriate statistical power for

enrichment analysis. Thirty targets is also the suggested minimum for motif discovery using DME (Smith, Sumazin et al. 2005). We assembled promoter sets for each of the 70 TFs using targets predicted by ARACNe, co-expression, and co-expression*, and identified enriched TRANSFAC motifs in each of the (70 x 3) sets. We refer to the ARACNe-inferred promoter set as the *conservation-free set* because it is assembled without regard to cross-species conservation. The co-expression* set was identified by taking the top *n* most-co-expressed genes, where *n* was the total number of targets identified by ARACNe rather than based on a predefined p-value threshold (see Co-expression in **Materials and Methods**). Note that there is no statistical threshold that could be used to reproduce the same selection a priori. Hence the co-expression* set can only be defined once ARACNe has been run and it was used only to determine if ARACNe further improves over co-expression even if only the most co-expressed targets are considered.

For each TF, the single most enriched motif was compared to the TRANSFAC reported motifs for that TF. Success was reported if a match was found using matcompare (Schones, Sumazin et al. 2005). The recall rate for conservation-free, co-expression, and co-expression*, was 27/70 (39%), 13/70 (19%), and 25/70 (36%), respectively (**Table 3-2**). In our experiments, ARACNe (conservation-free set) significantly outperformed co-expression ($p < 0.05$, by FET), and more narrowly outperformed co-expression*. This result suggests that ARACNe significantly improves over co-expression approaches by removing some false positives among high-scoring co-expressed targets and many false positives among low-confidence co-expressed targets.

Overall, the conservation-free set consistently had the highest recall rate, and its inferred targets were used for all subsequent experiments. Note that selection of the single most enriched motif for estimating recall is an exceedingly strict criterion. Indeed, we show that ubiquitous co-factors may in fact be more enriched than the TF-specific binding motif itself. As a result, the correct motif can be recovered for much more than 39% of the TFs if additional, statistically significant motifs, are also considered. For instance, when the criteria for correct identification was expanded to include the top 5 motifs (see Motif evaluation and discovery in **Materials and Methods**), recall improved to 48/70 (68%).

Table 3-2. Motif predictions comparison

	total TFs	**True Positives (matched TFs)**	**Recall**
ARACNe	70	27	38.57%
Coexpression	70	13	18.57%
Coexpression*	70	25	35.71%
ARACNe-MRC	70	25	35.71%

We compared motif enrichment accuracy across promoters corresponding to targets identified by the regulatory-network reverse-engineering algorithm ARACNe, co-expression, and a combination of both methods (see Results). ARACNe-MRC corresponding to ARACNe inferred target promoters with exons and repeats masked; see Table S1 for expanded description.

3.3.2 Cross-species conservation analysis further improves TFBS discovery

Many functional elements, including TFBSs, are conserved across species (Blanchette and Tompa 2002; Moses, Chiang et al. 2004), but the proportion of TFBS conservation that can be identified directly from genome alignments is still unknown.

Ward and Bussemaker (Ward and Bussemaker 2008) and Xie et al. (Xie, Cai et al. 2008) suggested using both alignment-based and alignment-independent approaches to identify evolutionary conserved regions. We studied the benefit of cross-species conservation in ARACNe-identified promoters for 70 representative TFs. Analysis was performed by sequence alignment, by pattern discovery using SPLASH (Hart, Royyuru et al. 2000; Sosinsky, Honig et al. 2007), and by a combination of the two methods. Since pattern discovery is especially sensitive to the presence of repeats and large highly-conserved regions, we first masked out repeats and coding exons. We processed the sequence data using the same procedure as described above to assess the recall rate for the known TFBS of the 70 representative TFs.

After masking repeats and coding exons, but before conservation analysis, the recall rate was slightly reduced, from 27/70 (39%) to 25/70 (36%), due to loss of some bona-fide TFBSs in masked regions (see **Tables 3-2**). However, this loss is requires for conservation analysis and is justified by the benefit of cross-species conservation. Additionally, the affected motifs for the two TFs were still ranked in the top five. In order to study the benefit of alignment-based conservation analysis, we used phastCons (Siepel, Bejerano et al. 2005) to identify the most conserved sequence fraction that would optimize recall (see **Table 3-3**), where this fraction is defined as the percent of nucleic acids in the sequences retained after masking poorly conserved regions. Surprisingly, the optimal recall rate using alignment-based conservation was only 25/70 (36%) at 10% DNA coverage, showing no improvement. We supplemented the alignment-based cross-species conservation with pattern-discovery-based (alignment-free) analysis.

Specifically, we first identified conserved patterns in each masked orthologous promoter set with SPLASH. We then selected the sequence fragments covered by the most statistically significant SPLASH patterns until the desired DNA coverage was achieved. This was also set to 10% to ensure results that are comparable with alignment-based conservation analysis. We refer to the resulting promoter fragment sets produced by the combination of phastCons and SPLASH analysis as the *combined-conservation set*. Analysis of the combined-conservation set improved the prediction recall rate to 31/70 (44%). Finally, we merged motif enrichment results independently produced by the conservation-free and the combined-conservation sets, re-ranking motifs according to the best classification relative-error rate achieved in either test (see **Figure 3-2**). The resulting recall rate increased further to 35/70 (50%). Thus, use of cross-species conservation data, within an integrative framework significantly ($p < 0.05$, by FET) improved recall rate, and joint use of alignment- and pattern-discovery-based approaches yielded an additional statistically significantly improvement ($p < 0.05$, by FET) over either method in isolation.

Table 3-3. Motif predictions based on conservation-free and conserved promoters

	total TFs	**True Positives (matched TFs)**	**Recall**
Conservation-free	70	27	38.57%
Alignment-based Conservation			
5%	70	7	10%
10%	70	25	35.71%
20%	70	22	31.43%

25%	70	18	25.71%
Combined-conservation			
10%	70	31	44.29%
Conservation-free and Combined-conservation merged	70	38	54.29%

All promoters used in these predictions were inferred by ARACNe algorithm. We compared motif enrichment accuracy across the original promoters and conserved regions identified using alignment-based conservation with varying DNA-coverage proportions, and a combination of alignment-based and pattern-discovery based conservation. For alignment-based conservation, best performance was achieved at 10% DNA coverage, and this was used in conjunction with pattern-discovery based conservation at 10% DNA coverage to produce combined-conservation. A test is considered as successful if the most enriched motif identified using either the conservation-free or the combined-conservation promoters matched the known motif for the TF. We called it conservation-free and combined-conservation merged. The recall rate at this level was significantly better than that using conservation-free alone.

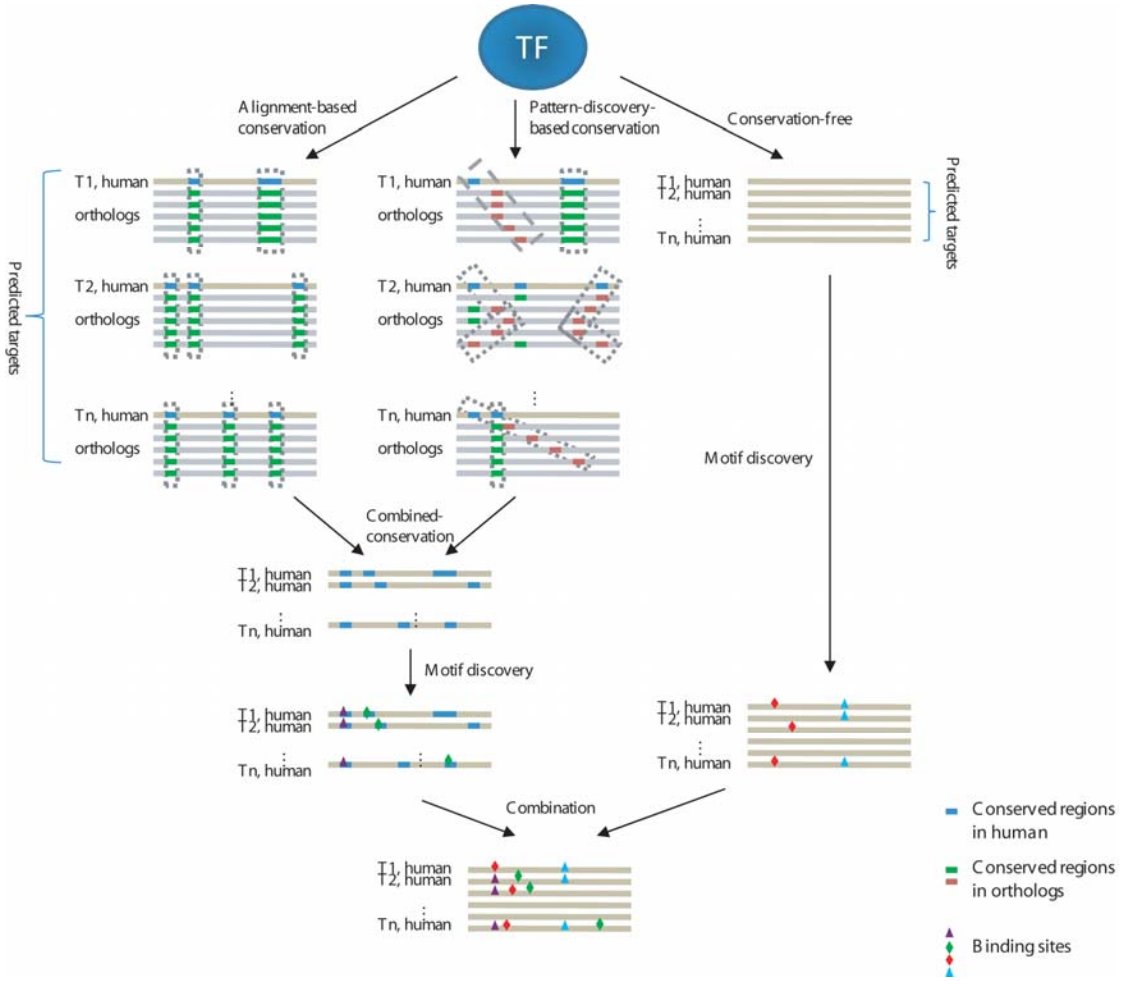


Figure 3-2. OmniMiner's motif discovery workflow. For each TF, we identified target genes for the TFs and assembled a set of promoters corresponding to these genes. Cross-species conserved regions were identified in these promoters using alignment-based and pattern-discovery-based methods and were combined to produce the combined-conservation set. Motif discovery was performed separately on the original promoters and the combined-conservation set. The resulting motifs were merged and re-ranked according to their classification relative-error rate.

3.3.3 Testing *de novo* motif discovery

To estimate OmniMiner's accuracy and determine if our test set has is rich enough for *de novo* motif discovery, we tested whether TFBSs for the 38 TF, whose TRANSFAC motifs were correctly identified in the previous subsection (on either conservation-free or combined-conservation sets) could also be identified *de novo*. We also used the analysis of the combined-conservation set to compare the performance of OmniMiner to that of GibbsModule. Specifically, we ran DME (Smith, Sumazin et al. 2007) on both the conservation-free and combined-conservation sets and recorded p-values for DME-identified motifs, reporting motifs with $p < 0.05$ (see Motif evaluation and discovery in **Materials and Methods**). Following the same procedure described for TRANSFAC motifs, we re-ranked significant motifs based on the best classification relative error rate achieved on either the conservation-free or the combined-conservation sets. We considered DME to be successful if a top *de novo* discovered motif matched a known motif for the TF according to matcompare. Results are given in **Figure 3-3**. For 32/38 (84%) of the TFs, we were able to discover significantly enriched motifs that

matched the reported TF motif in TRANSFAC. Strictly matching significant motifs among the top 5 motifs per TF were recovered for 2/38, 13/38 and 10/38 of the TFs on the conservation-free, combined-conservation, and the combination of the two, respectively. The result suggests that, likely because of their length and count, cross-species conservation is required for *de novo* discovery on our promoter test sets.

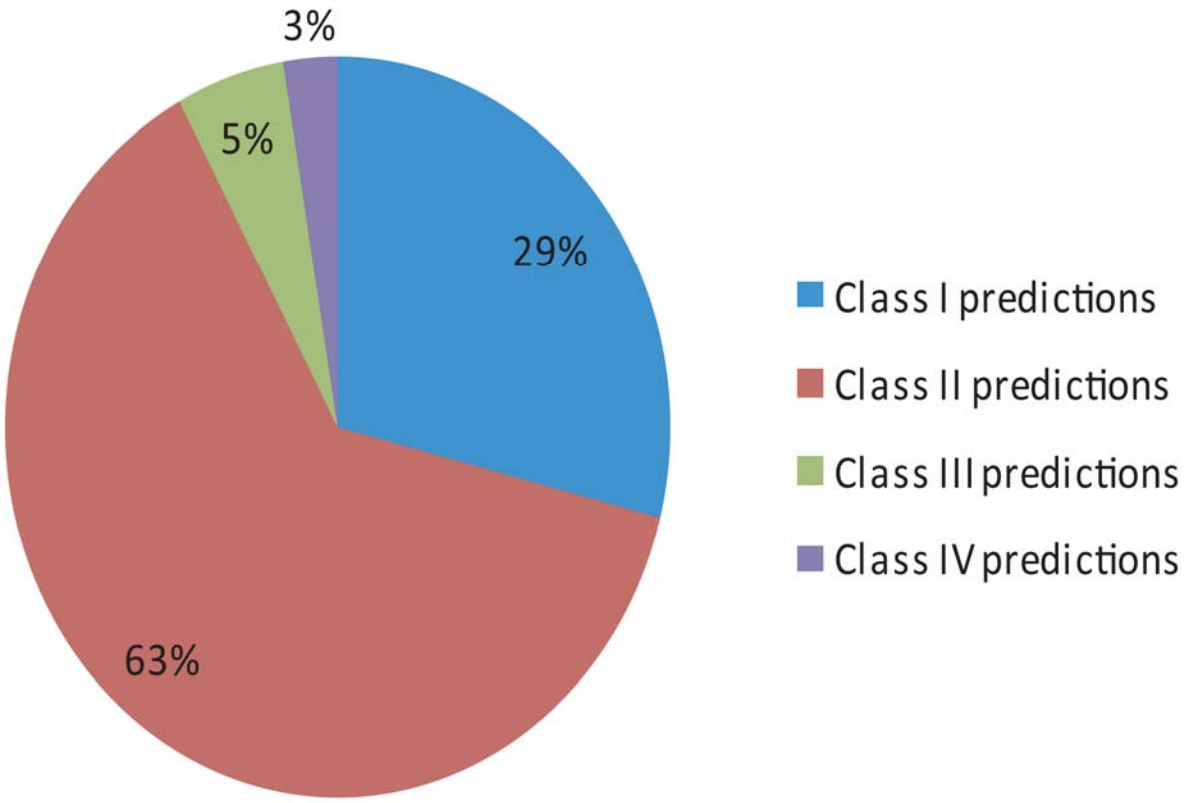


Figure 3-3. *De novo* motif-discovery accuracy measurement. *De novo* motif-discovery was performed on the 38 TFs for which the known TFBSs were enriched in the target genes. Predicted motifs were classified into four classes. Class I: the top three predictions included a significant classifying motif than matched the known motif for the TF. Class II: a lower-ranking significant classifying motif that matched the known motif

for the TF. Class III: The most enriched motif was a significant classifier, but no motifs matching the known motif for the TF. Class IV: no significant classifiers were found.

In order to compare OmniMiner on combined-conservation promoters to GibbsModule, we ran GibbsModule on the conservation-free *set* with the orthologous promoters as additional input. GibbsModule performs cross-species conservation analysis internally, but it does not output p-value information or motif ranking. In the absence of ranking, we used all GibbsModule-discovered motifs, and for fairness, compared the 9 GibbModule-discovered motifs both to the top 3 and to the top 9 DME-discovered motifs with no p-value restriction. For 12/38 (31%) of the TFs, one of the nine GibbsModule-discovered motifs matched a known motif for the TF. This performance is significantly worse than DME's recall rate of 21/38 (55%) when considering the top nine ranking motifs, and it is also worse than the recall rate of 14/38 (37%) when only the top three DME motifs are considered (see **Table 3-4**).

Table 3-4. Performance comparison of OmniMiner to GibbsModule

	total TFs	**True Positives (matched TFs)**	**Recall**
DME-Total			
(top 3)	38	11	28.95%
(top 3)*	38	14	36.84%
(top 9)	38	17	44.74%
(top 9)*	38	21	55.26%
GibbsModule			

(best 9)	38	12	31.58%

We compared OmniMiner and GibbsModule recalls on our 38 TFs test set.

DME-Total used both the conservation-free and the combined-conservation promoter sets.

* No p-value threshold was used for pruning motifs.

To test DME-discovered motifs on an external data source, we tested for site enrichment for the top three predicted ZNF263 motifs in the top 500 ZNF263-bound regions according to ChIP-seq in K562 cells relative to unbound regions after size correction. The results show that all top motifs are highly enriched in ZNF263-bound regions with respective p-values of 2.35e-57, 3.80e-19 and 9.01e-3 (see Validation in **Materials and Methods**).

3.3.5 Test set clusters

We clustered test sets and motif discovery methods according to motif discovery success; see **Figure 3-4**. The clustering results show clear TF grouping according to binding site identification and discovery methods, suggesting that, for some TFs, binding site enrichment can be detected using most methods. However, for some TFs, TFBS enrichment is only detectable when cross-species conservation data is used. For example, STAT1, STAT2, STAT4, STAT3, RELA, MAF, MYC and IRF7, which form one cluster, were correctly classified with and without conservation analysis, and using known motifs or *de novo* discovered motifs. Members of a second cluster, including PAX9, POU2F1, CEBPA, MYB, PAX8, E2F1, ARNT, and AHR1, were correctly classified with and without conservation but not using *de novo* motif discovery. Finally,

members of a third cluster, including JUND, ETS1, ZNF42, SMAD2, LEF1, TAL1, FOXC1, TGIF, and SMAD1, were correctly classified with the help of cross-species conservation but not in the original conservation-free promoter sets.

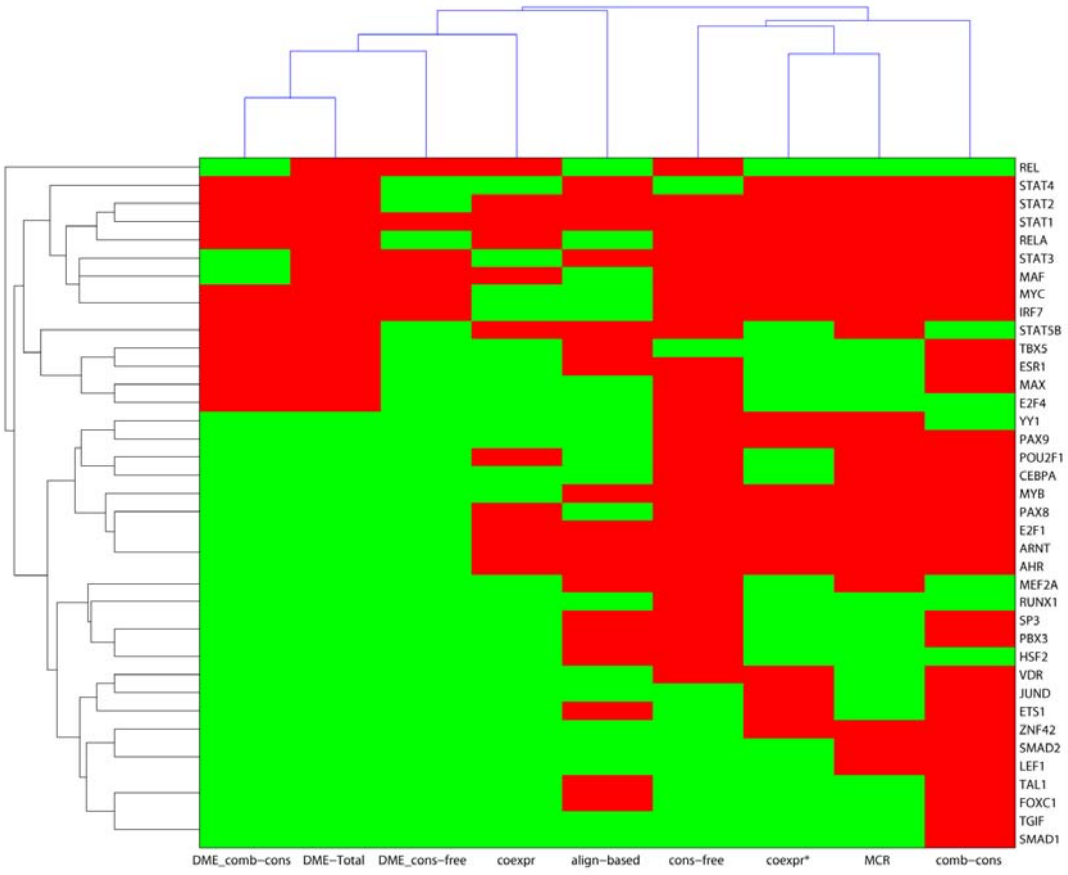


Figure 3-4. **Classification of motif prediction.** We classified the 38 TFs used as the test set and the motif discovery methods according to enrichment and discovery success. DME_comb-cons stands for the comb-cons promoter set that was used for *de novo* motif discovery. DME-Total represents the result from combining *de novo* motifs discovered in the conservation-free and combined-conservation sets. DME_cons-free stands for the cons-free promoter set that were used for *de novo* motif discovery. Coexpr represents the promoter set inferred by Spearman correlation. Align-based represents the promoter set in

which the conservations were identified by alignment-based method. Cons-free represents the conservation-free set. Coexpr* represents the promoter set inferred by the combination of the ARACNe and coexpression. MCR represent the conservation-free set with exons and repeats masked. Comb-cons stands for the promoter set in which the conservations were identified either by a combination of alignment-based and pattern-discovery-based methods.

3.3.6 New predictions and biochemical validation

The TRANSFAC E2F1 DNA-binding motif M00918 was the most enriched motif with sites in promoters of predicted E2F1 targets. As proof of principle, we tested top-scoring sites for M00918 in seven randomly-selected promoters using quantitative PCR of chromatin immunoprecipitation assays (qChIP). Our results show that E2F1 binds to top predicted sites in the promoters of RAD54L, MCM2, PKMYT1, FANCG and GTSE1. We failed to show binding to top sites in the promoters of SAC3D1 and PPM1G (**Figure 3-5A**).

No motif was significantly enriched in the JUND conservation-free *promoter sets*. However, both JUN and E2F-1 motifs (M00428 and M000172) were among the top 5 motifs for all JUND test sets, and they were significantly enriched in the JUND combined-conservation set, where the reported AP1 and E2F1 motifs were the most enriched motifs. Consistent with these predictions, we show that both E2F1 and JUND bind to the best matching sites in the promoters of predicted targets C21orf2, PTDSR, KIAA0556 and PER1, suggesting that this transcriptional co-binding pattern may be

pervasively used in human B cells. However, while we were able to show that JUND binds to RPS29 and RPS25, we could not detect enrichment of E2F1 antibody by qChIP at the top candidate sites in their promoters. Similarly, we failed to show enrichment of either TF's antibody to the promoter of the predicted target DYRK1B (**Figure 3-5B**). This suggests that either the qChIP analysis produced a false negative result or that these sites were false positive predictions. In total, we validated the top predicted sites for JUND binding in 6/7 of the predicted JUND targets, and we validated E2F1 binding sites in 4/6 of the JUND bound promoters.

Finally, we proceeded to predict *de novo* DNA-binding motifs for twenty TFs with previously uncharacterized binding motifs. The top three predicted motifs with p-value < 0.05 (see Motif evaluation and discovery in **Materials and Methods**) for each TF are given in **Appendix I**; only two significant motifs were identified for NME2 and EP300. Experiments on our test set suggest that *de novo* motif discovery is able to identify significant binding motifs for the vast majority of TFs. Because of antibody availability, we chose to validate binding sites for the top BCL6-predicted motif. Our results show that BCL6 binds to the promoters of LRMP, MKI67, RGS13, STK39, BCL7, H2AFX and VGLL4 (**Figure 3-5C**). Indeed, all tested promoters were validated for BCL6 binding. The BCL6 motif was identified from the BCL6 combined-conservation set, and matched a previously reported BCL6 half site (Kawamata, Miki et al. 1994).

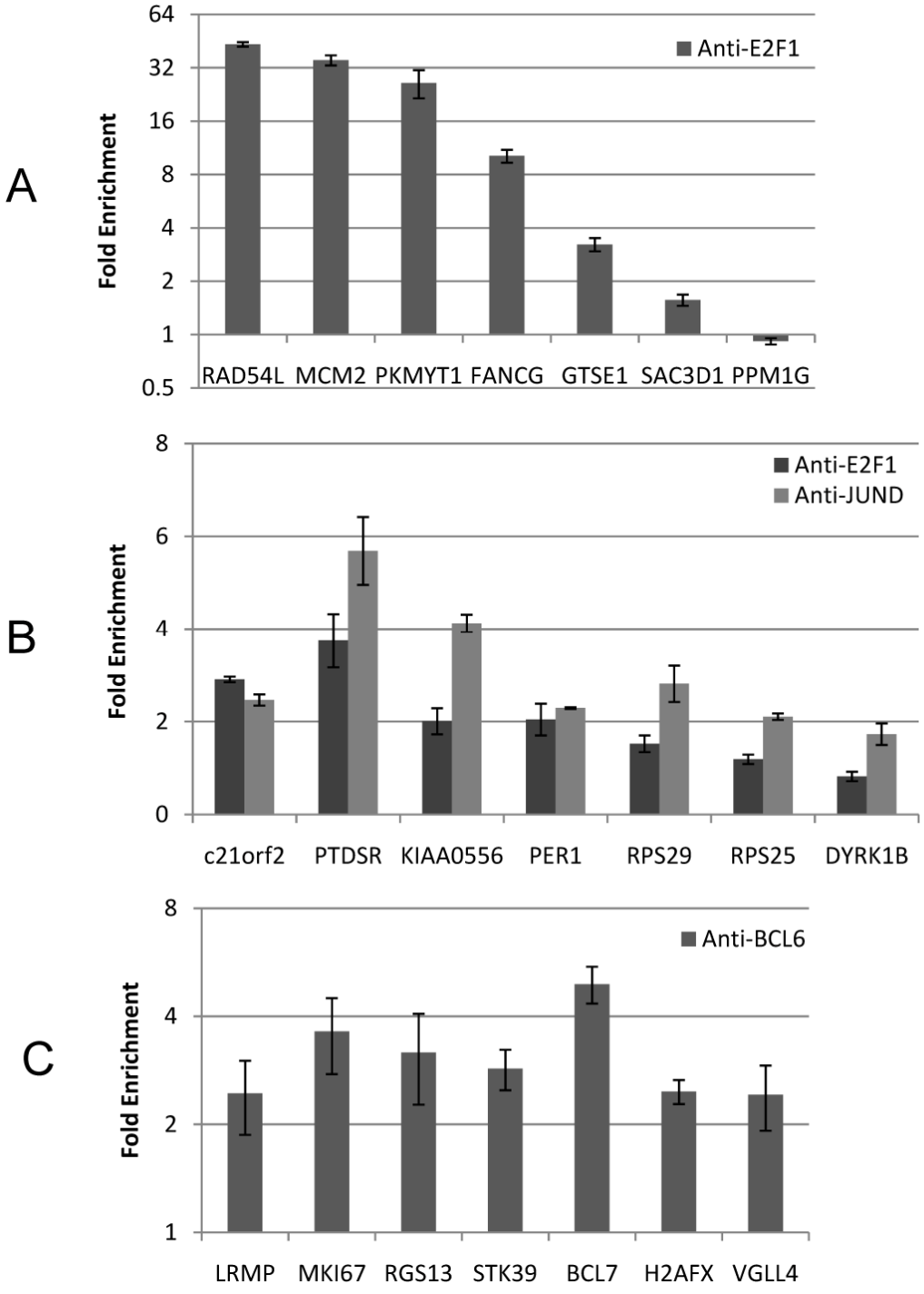


Figure 3-5. Binding validation. (A) E2F1 binding to predicted E2F1 targets. **(B)** E2F1 and JUND binding to predicted JUND targets. **(C)** BCL6 binding to predicted BCL6 target. We plot fold enrichment relative to IgG (mean ± s.e.m)

3.3 Discussion

Here, we proposed a novel integrative methodology that combines reverse-engineering of transcriptional networks and cross-species conservation analysis for TF binding-motif discovery. We produced *de novo* motif predictions for 20 previously uncharacterized TFs, and validated site predictions for the top BCL6 motif and for co-binding patterns between E2F1 and JUND. In order to compare methods, we produced an extensive test set of co-regulated human genes and promoters in B cells; these test sets are given in Table S3. Our results suggest that 50% of the transcriptional regulators analyzed by ARACNe in a human B cell context produce inferred target sets that are significantly enriched in their bona-fide TF functional direct targets. This is a lower-bound for the proportion of TFs for which bona-fide targets can be identified since (a) only a relatively small region of the promoter was considered, (b) some TFs are poorly characterized in TRANSFAC, (c) only predicted activated targets were considered, (d) some TFBS motifs are highly degenerate and may not be reconstructed from enrichment analysis alone, and (e) we defined success restrictively, requiring the top predicted motif to match a known motif to the TF and disregarding the possibility of a match to a co-factor motif.

The novelty in our approach was three-fold. First, we showed that using a reverse-engineering algorithm instead of gene co-expression to identify TFBS-enriched promoter sets significantly improved prediction. Second, we showed that using a combination of alignment- and pattern-discovery-based conservation analysis approaches significantly improves prediction when compared to using only one of the approaches.

Third, we showed that by combining the two approaches, we can further improve prediction accuracy and almost doubled the 12/38 (31%) recall of another recent integrative approach (GibbsModule). Finally, we produced predictions for 20 TFs with previously unknown binding affinity and validated predictions by quantitative Chromatin Immunoprecipitation assays (qChIP) and enrichment in ChIP-seq data. By developing a unique test set of human promoters and conserved regulatory regions, we were able to produce realistic estimates for the quality of our *de novo* prediction method.

We used stringent criteria to test our input sets, requiring that the most enriched TRANSFAC motif in a foreground set be similar to a known motif for the TF. Based on this metric, even before cross-species conservation is used, nearly 40% of the tested TF motifs pass this criterion. This is a significantly higher recall rate than the one observed when using co-expression. To understand the source of this performance gap, we compared the two methods to a hybrid method. Instead of using a p-value cutoff for co-expression, we used the number of ARACNe predicted activated targets as a cutoff. The hybrid method performed only marginally worse than ARACNe, suggesting that genes with expression profiles that are most similar to the TF's are the most likely targets and that ARACNe's main advantage is in a highly TF-dependent and accurate co-expression similarity cutoff selection.

Maybe the most surprising result in this study was that using alignment-based conservation to identify regions enriched with TFBSs did not improve recall. This

suggests that removing less conserved regions, in an effort to reduce background noise, may lead to loss of regions containing bona-fide TF biding sites. On the other hand, using non-linear pattern-discovery-based conservation improved the performance considerably and use of both methods in combination provided the best results. Cross-species conservation significantly improved recall, but only when jointly considering sequence fragments discovered both with alignment-based and pattern-discovery-based analysis. For seven TFs, known motifs were found to be the most enriched only when using the entire promoter sequence, suggesting that evolution of their transcriptional targets may be more recent and poorly conserved in orthologous species or that our alignment techniques are not sensitive enough for this task. Our conservation-free promoter sets appear to be either too long or too few for *de novo* discovery, which was successful only on the combined-conservation set, and for 18/20 of the TFs for which we discovered *de novo* DNA-binding motifs, including the validated BCL6, the top motifs were selected from the combined-conservation set analysis.

Figure 3-4 suggests that our test sets can be clustered according to motif discovery success, with one 8-test-set cluster consisting of promoters that were correctly classified without conservation, with the aid of alignment-free conservation, and by *de novo* motif discovery. However, only 4/8 of the sets were correctly classified using alignment-based cross-species conservation. Our findings support the idea that TFBS conservation is fundamentally different from coding-region conservation. This may be due to operating distance flexibility, *cis*-regulatory module grammar, or neutral mutations in site positions that correspond to low-information motif columns.

Despite these significant advances, we could not identify known TFBS motifs for several of the TFs, suggesting that these are either poorly characterized in TRANSFAC, that binding for that promoter is supported by heterogeneous mechanisms, or that reverse-engineering may fail to appropriately characterize the transcriptional targets of some TFs. This, in turn, affirms that the problem of TF binding-site characterization is still open and much remains unknown. It also suggests a set of TFs that may be especially hard to characterize. An important point is that our ability to characterize TF binding motifs is likely cell-context dependent. We used a large gene expression profile dataset for mature human B cells, which may have both improved our ability to characterized some TFs as well as hinder the ability to characterize others. Analyses of similar datasets from other cellular contexts may help answer these questions.

Machine learning heuristics fall in one of three categories: heuristics that search for good solutions in complete problem domains but do not guarantee optimality, heuristics that discover the best solutions in simplified problem domains, and those that search in simplified problem domains but do not guarantee optimality. GibbsModule arguably falls in the first category, while DME, SPLASH and OmniMiner belong to the second category. We previously showed that DME outperforms other motif discovery algorithms on both synthetic and mammalian data. The argument in favor of DME (Smith, Sumazin et al. 2005) is based on properties of the motif-discovery solution-space structure, which under a variety of formulations is in $\Omega(n^m)$ and $O(2^{nm})$, where m

denotes the number of input sequences and *n* denotes their length. This space is smooth and allows for local optima discovery, making DME's fine grid search followed by a locally optimal refinement a successful strategy. In the presence of orthologous promoters, the search space is in $\Omega(n^{dm})$, where *d* is the number of ortholog species used. Moreover, there is no proven formulation for the integration of the two orthogonal optimization criteria. We hypothesize that due to the computationally prohibitive task of identifying patterns across sequences with varying degrees of similarity and in the absence of a demonstrably good type-1 method, a type-2 heuristic should be preferred. Finally, our success in identifying pattern-discovery-based conserved regions is due to SPLASH's ability to identify long and sparsely conserved regions. Thus, SPALSH is able to overcome some of the limitations of linear multiple-sequence aligners, and specifically it does not discard sites due to varying module grammar, or neutral mutations. We followed SPLASH conserved-region identification with motif discovery by DME to identify conserved motifs in these regions, thus fixing motif column values whether they have high or low information content.

To create a realistic testing platform for motif discovery in human regulatory regions, we identified promoter sets that were predicted to be co-regulated by known TFs, and are significantly enriched with a motif associated with these TF. This platform allowed us to estimate the accuracy of our motif discovery methods. The platform is composed of 38 human promoter sets of varying sizes and it is computationally validated. Its size, validation, and specialization make it a unique platform for motif discovery evaluation. Our tests with *de novo* motif discovery suggest that we recover 12/38 of the

known motifs associated with the query TF, and we identify significant motifs for 36/38 ($p<0.01$) of these TFs.

Chapter 4

coEDGi, a TF-centric enhancer discovery approach

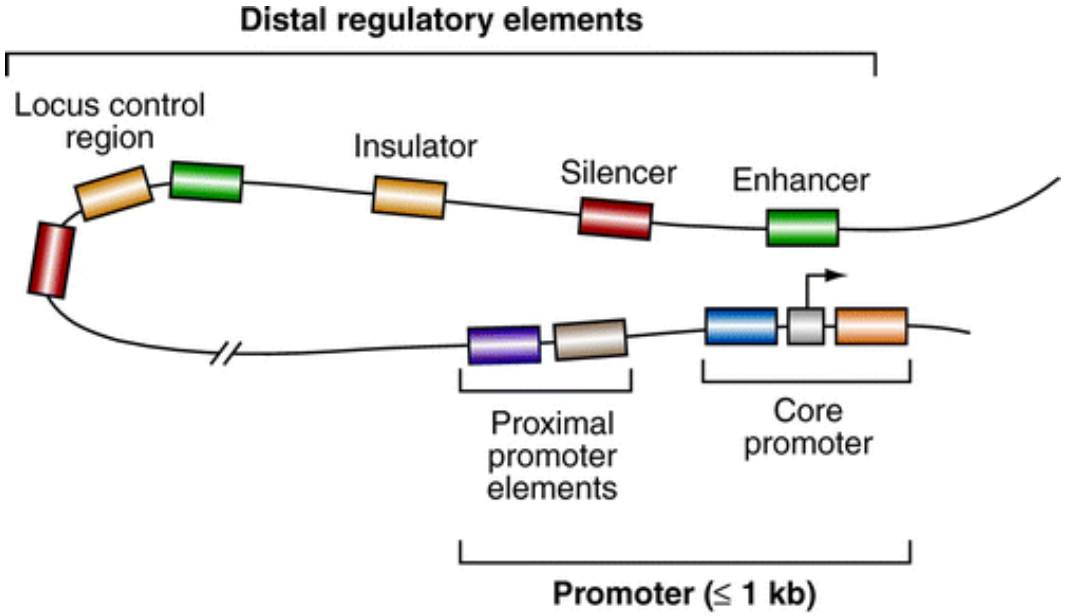
--- integration of gene co-regulated and genomic information

4.1 Introduction

We live in a complex society where we interact with many different people in our daily lives. Inside our body, different organisms work together to ensure we perform normally day and night. This mechanism of different parts working together is observed in gene expression regulation. Like the actors in an orchestra, different TFs bind to different sites of target genes at scheduled times to turn expression of gene on and off. In addition, by combining with different partners, the same TF can perform different functions in different cell types as well as on different target genes. In **Chapter 3**, our study was focused on single TF. We developed a strategy to *de novo* predict TF binding motif for a given cellular context. In this chapter, we extended to understand how TFs worked together by identify their working unit, '*cis*-regulatory module (CRM)'.

Genome complexity is not defined by the number of genes the genome contains, but the number of CRMs in the genome. In another word, how genes are regulated. For example, although the *Drosophila melanogaster* (fruit fly) genome contains less genes

(Adams, Celniker et al. 2000) than the *Caenorhabdotis elegans* (worm) genome (Waterston and Sulston 1995) (14,000 *vs.* 20,000), *D. melanogaster* genome complexity is much higher than that of *C. elegans*'. This is because on average, fly gene is regulated by three or four different CRMs, while worm gene only has one or two. CRMs can be categorized into several types, including enhancers (Banerji, Rusconi et al. 1981), silencers (Brand, Breeden et al. 1985) and insulators (Bell, West et al. 2001) (**Figure 4-1**). In our study, we only focused on enhancers (CRMs and enhancers were used interchangeable in the rest of the paper). Enhancers have been widely studied in different model organisms, such as flies, worms, sea urchins, mice and human. Let's first take a close look of the structure of enhancer. A typical enhancer is usually 300bp to 1kb in length and mediates gene expressions in specific cell context (Arnone and Davidson 1997). Insider enhancers, there are multiple binding sites for both transcriptional activators and repressors (Markstein and Levine 2002). In contrary to typical TF binding sites that are near the gene promoter regions, enhancers can be found all over the genome. It could either be close to or several kilo base pairs away from the transcription start site (TSS) of its target gene. Its location is not limited to the upstream promoters, downstream or even introns could also harbor it (Blackwood and Kadonaga 1998). The wide range of possible location sites make enhancer hard to detect.



Maston GA, et al. 2006.
Annu. Rev. Genomics Hum. Genet. 7:29–59

Figure 4-1. Schematic of a typical gene regulatory region.

The promoter, which is composed of a core promoter and proximal promoter elements, typically spans less than 1 kb pairs. Distal (upstream) regulatory elements, which can include enhancers, silencers, insulators, and locus control regions, can be located up to 1 Mb pairs from the promoter. These distal elements may contact the core promoter or proximal promoter through a mechanism that involves looping out the intervening DNA (Maston, Evans et al. 2006).

Experimental approaches to identify enhancer regions are beyond the scope of this paper_and are described by Elnitski (Elnitski, Jin et al. 2006; Maston, Evans et al. 2006). In the last several years, a large number of computational methods have been

developed to detect and predict enhancer regions (Blanchette, Kent et al. 2004; Elnitski, King et al. 2006; Brown 2008; Schultheiss, Busch et al. 2009). But these methods have never been evaluated together like Tompa *et al* did in 2005 (Tompa, Li et al. 2005). In addition, most of these methods have only been applied to lower species, but not higher species, such as fruit fly. We also noticed that, a lot of pre- and pro- processes are required for those methods in order to get meaningful results. Based on the data inputs of those methods, they can be categorized into three groups. Methods in the first group tend to use a prior knowledge, such as genes were from the same pathway or co-regulated. The assumption is functionally related genes are likely be regulated by the same set of enhancers. Methods in the second group were more favoring comparative genomic, also known as phylogenetic footprinting (Tagle, Koop et al. 1988). The assumption is that enhancers are functionally important elements and thus are likely to be highly conserved through evolution. But it has also been shown that highly conserved sequences might have any known types of function (Cooper, Stone et al. 2005; Siepel, Bejerano et al. 2005) and some functional important modules only conserved in closely related species. Therefore, a variation called phylogenetic shadowing (Boffelli, McAuliffe et al. 2003) was developed, which only compares closely related sequences. Methods in the last group absorbed the advantages of previous two groups by integrating both. They started with a gene set that were selected with a prior knowledge and used evolutionary information of each individual gene to identify functional important regions. The predictions were based on the common functional important regions across all genes in the set. Our method belongs to the third group.

Currently, CRM discovery algorithms have a number of limitations. Some of them started with a pre-defined enhancer module and scan the promoter regions for potential matched sites (Erives and Levine 2004). Others scan the promoter regions for known TFBSs and looked for clusters that are enriched of these TFBS (Sharan, Ovcharenko et al. 2003; Schroeder, Pearce et al. 2004). A prerequisite for these methods is either a robust enhancer module or well-defined TF motif profiles. Unfortunately, building a robust enhancer module requires a lot of prior knowledge and many TFs don't have well characterized motifs. Another limitation of those methods is that only a small region close to the TSS was covered. For example, CRÈME algorithm only retrieved 1200bp upstream of TSS for each gene (Sharan, Ovcharenko et al. 2003). As discussed above, enhancers are widely spread. Narrowing the search space only at proximal promoter regions will result in missing a lot of potential enhancer sites. Although there are some algorithms capable of identifying highly conserved regions from a long range (Brown 2008), only one gene can be analyzed at a time.

In this study, we introduced a new enhancer discovery algorithm, coEDGi (*co*-regulated based Enhancer Detecting using Genomic information), that enables users to maximally detect potential enhancer sites. There are two innovations in coEDGi. First, it is a TF-centric enhancer detecting approach. The predicted enhancers are directly linked to the TF of interest. Second, due to the implementation of a non-alignment based pattern discover approach, searching long ranges of genome for maximal retrieval of potential enhancer sites became possible.

A Hox protein Sex combs reduced (Scr) was selected for our study. To detect the enhancers that might harbor Scr binding sites, we started with a set of co-expressed genes that were regulated by Scr (**Table 4-1**). Previously study showed that together with the PBC factor Extradenticle (Exd), Scr-Exd specifically bound to forkehead gene (fkh) in *D. melanogaster* and activated fhk gene expression (Ryoo and Mann 1999; Andrew, Henderson et al. 2000). Therefore, even though fkh gene was not co-expressed with those target genes, it was added into the gene set as a positive control. We implemented comparative genomics by selecting orthologous sequences for from 12 fly genomes to identify highly evolutionarily conserved regions for each target gene. And for each gene, its upstream, intron and downstream regions were examined. Although highly conserved, not all regions were similar to each. We hypothesized that if the conserved regions shared common Scr binding sites, they should be more similar to each other. Based on this, evolutionarily conserved regions that shared significant amount of common patterns were used at the pool for enhancer discovery. *De novo* motif discovery algorithm was applied to predict and locate most enriched binding motifs in the selected sequence pool. Regions that enriched of binding motif clusters were selected and evaluated. We show that our coEDGi correctly detected the known enhancer site for fkh gene. The workflow is demonstrated in **Figure 4-2**.

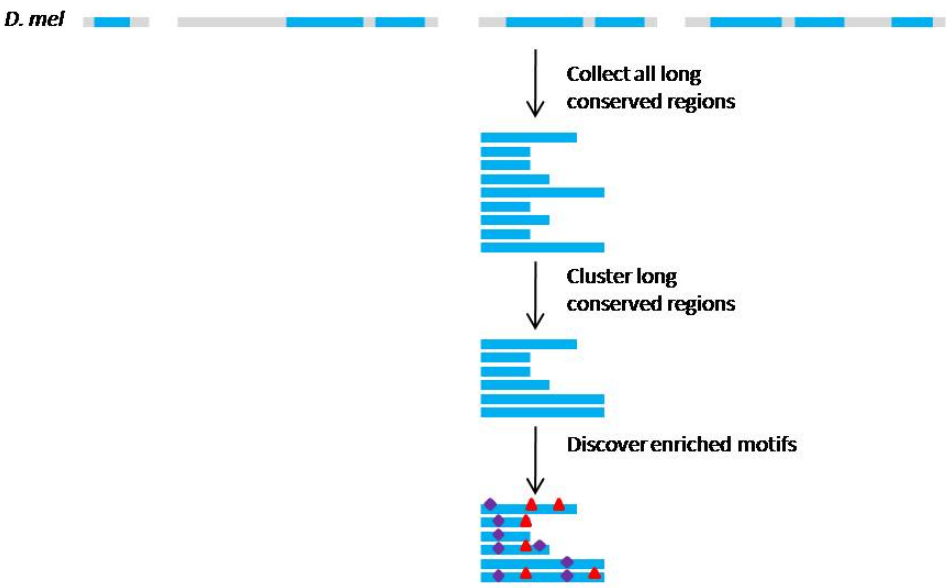
Table 4-1. Potential Scr regulated genes in *D. melanogaster*

FLYBASE ID	ANNOTATION SYMBOL	SYMBOL	

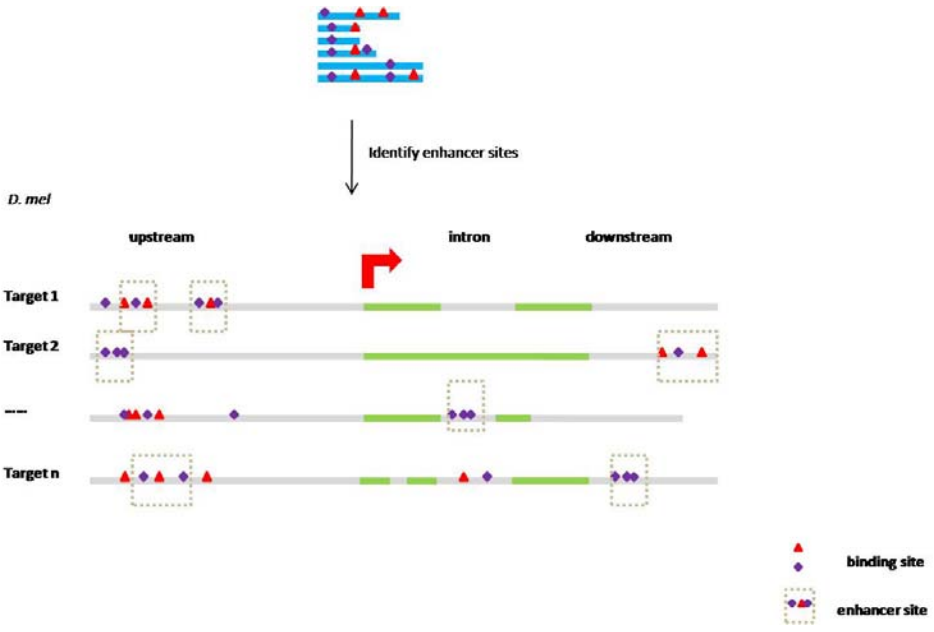
FBgn0020389	CG8363	Papss	Upregulated by Scr, downregulated by other Hox proteins	Likely expressed in salivary gland	Known to be expressed in salivary gland	Involved in sulfation pathway
FBgn0038524	CG7623	sll				
FBgn0030498	CG32632	Tango13				
FBgn0003089	CG9614	pip				
FBgn0039779	CG1546	PH4alphaSG2				
FBgn0037672	CG12952	sage				
FBgn0036470	CG13463	CG13463				
FBgn0036167	CG33272	CG33272				
FBgn0039682	CG7584	Obp99c				
FBgn0036469	CG18649	CG18649				
FBgn0036390	CG13738	CG13738				
FBgn0037179	CG14453	CG14453				
FBgn0039098	CG13822	CG13822				


TF
Target gene
Get noncoding regions
upstream
intron
downstream
D. mel
11 orthologs
Discover conserved sites
Cluster conserved sites
Collect all long conserved regions
Cluster long conserved regions
Discover enriched motifs
Identify enhancer sites
Target 1
Target 1
Target n
TSS
Noncoding region
exon
short conserved site
long conserved region
binding site
enhancer site
gene
upstream
intron
downstream
D. mel
11 Orthologs

a.



b.



c.

Figure 4-2. Workflow of coEDGi algorithm

4.2 Materials and Methods

4.2.1 Scr target gene list

Scr target genes were selected by Dr. Matthew Slattery from Professor Richard Mann's group. In total 13 genes were selected. They were Papss, sll, Tango13, pip, PH4alphaSG2, sage, CG13463, CG33272, Obp99c, CG18649, CG13738, CG14453 and CG13822. All the genes were up-regulated by Scr but down-regulated by other Hox proteins. Papss, sll, Tango13, pip, PH4alphaSG2, sage and CG13463 were known expressed in salivary gland. CG33272 was likely to express in salivary gland. Papss, sll, Tango13, pip were also involved in sulfation pathway. Another known Scr target gene fkh, although was not in the co-regulated gene list, was also added to the target gene set as a positive control because we knew its enhancer site.

4.2.2 Sequence retrieval procedure

For each target gene, we applied the same procedure to retrieve its sequence segments. First, gene symbol and Flybase id of the gene were mapped to the Refseq id based on the annotation file downloaded from Flybase, version 2009-02 (flybase.org). Second, upstream, intron and downstream sequence coordinates of the gene were calculated based on the gene coordinate associated with Refseq id. Sequences that corresponded to the coordinates were retrieved from Flybase's gene sequence database. Exons were masked in each sequence segment. Repeats were masked using Repeat Masker method obtained from institute of systems biology (http://www.repeatmasker.org/RMDownload.html). Standard parameters setting of

Repeat Masker were applied. Because some genes didn't contain any introns, no any intron sequence segments were generated from those genes. Third, up to 11 orthologous genes were used to retrieve orthologous sequences for the each *D. melanogaster* gene. These orthologous fly genomes were *D. pseudoobscura, D. sechellia, D. ananassae, D. erecta, D. grimshawi, D. yakuba, D. mojavensis, D. persimilis, D. simulans, D. virilis* and *D. willistoni.* The *D. melanogaster* gene's corresponding orthologous coordinate in each fly genome was obtained from Flybase's ortholog annotation file. Not all target genes have all 11 orthologous genes. Tango13 and CG33272 had no annotations in Flybase. Pip only had one annotated orthologous gene. All three of them were discarded from further analysis.

4.2.3 Detecting local permutation clusters (LPCs) with SPLASH and Promoclust

For each target gene, we applied the same procedure to detect highly evolutionarily conserved regions. Each target gene had a sequence set which contains all the orthologous sequences. SPLASH (Califano 2000) was applied to identify the conserved patterns (SCMs) in orthologous sequence set. The parameters setting was following 1) minimal pattern length was 8 nts and was increased to 15 nts maximal, 2) the minimal DNA conservation density (a.k.a. number of conserved nucleotides in the pattern) was 75% (i.e. 6 nts out of 8 nts) and was increased to 100%. Identified patterns were ranked and selected all patterns until the sum of the conserved nucleotides in the patterns added up 15% of the total sequence or 1500bp, whichever was longer. Selected SCMs were used as the inputs for Promoclust method (Sosinsky, Honig et al. 2007). Each

SCM was used as an initial LPC seed which can be extended by recursively combining with other SCMs until the predefined maximum cluster length (w) was reached. LPC was therefore a set of SCMs occurred in a w length window, whose boundaries were defined as the positions of the leftmost and the rightmost SCMs it contained. 1000bp was defined as the maximal cluster length in our study. Each LPC was only represented by one canonical SCM set and the same SCM was then used to search the occurrences of other possible LPCs in other regions of the sequences as well as in the orthologous sequences. When two LPCs were compared, we didn't take the order of the SCMs inside and the frequency of each SCM into consideration. As long as the two LPCs contained the same SCMs content, they were treated as a match. For instance, even though LPC1 contains {SCM1, SCM15, SCM23} and LPC2 contains {SCM23, SCM1, SCM23, SCM15}, they were still considered as the same LPC with {SCM1, SCM15, SCM23} as the canonicity. Every canonical LPC was exhaustively searched across all sequences. In our study, we required 1) at least two distinct SCMs in the identified LPC, 2) the maximal length of the LPC was 1000 bps, 3) the identified LPC must occurred in at least 50% of the input sequences. All the LPCs that met predefined criteria were reported. The conservation score of each LPC were calculated based on the total number of unique SCM nucleotides inside. The statistical significance of LPC was estimated and only LPCs with p-value <= 0.05 were kept for further analysis.

4.2.4 Clustering LPCs based on SCM similarity

To cluster LPCs, we compared the all LPC pairs. The similarity of two LPCs was based on the shared SCMs inside each LPC. As demonstrated in **Figure 4-3**, a 200bp window was selected from the query LPC and the LPC to be compared. SPLASH was used to retrieve common SCMs. Similarity score between the 200bp sequences from two LPCs was the sum of all unique nucleotides in the common SCMs. The p-value of the similarity score was calculated by compared to two random 200bp sequences. Similarity scores with p-value <= 0.05 was kept. The 200bp window moved along LPCs on at a 100bp/step. For each 200bp window, a best mapping score was achieved after compared to all possible 200bps from the other LPC. The similarity between two LPCs was the sum of all the best 200bp similarity scores. All LPCs were compared and their distance matrix was used in the following hierarchy clustering. Clustered LPCs subgroups were compared to randomly generated LPCs group to estimated statistical significance. Only subgroups with p-value <= 0.05 were kept for further analysis.

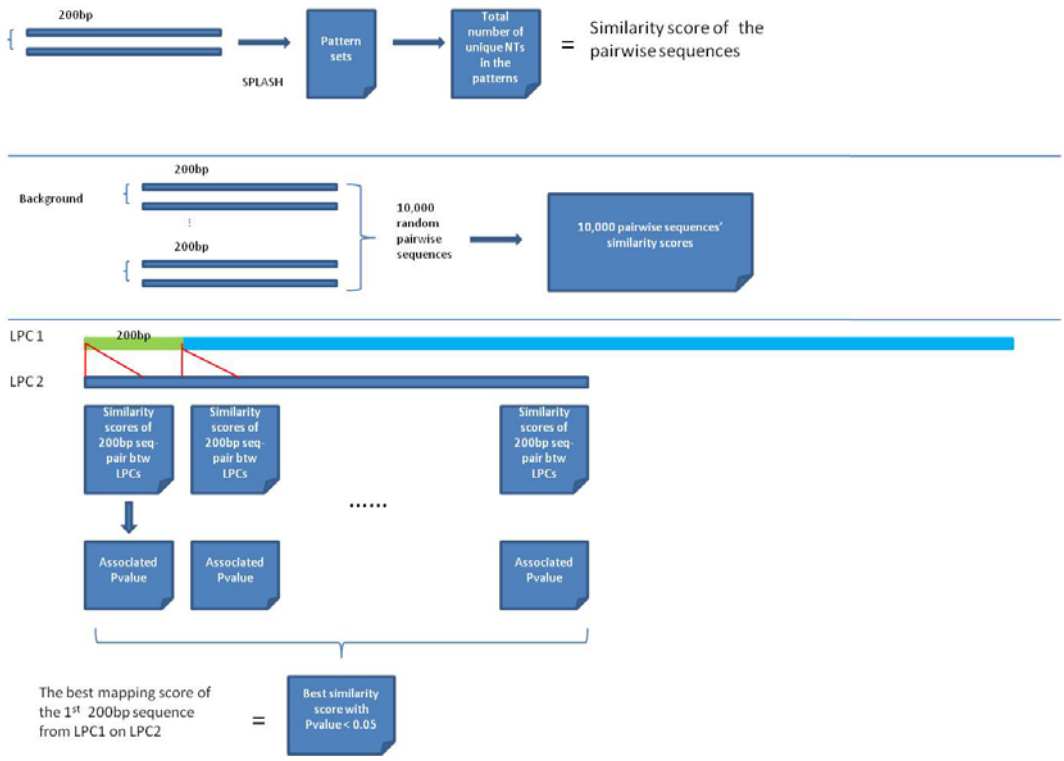


Figure 4-3. LPC similarity comparison strategy

4.2.5 Predict enriched motif with DME and cluster enriched motifs with Promoclust

DME was used for *de novo* prediction of the enriched motifs in the selected LPCs (Smith, Sumazin et al. 2007). The parameters setting was the same as described in **Chapter 3**. All predicted motifs' statistical significances were estimated by compared to motifs predicted from 1000 randomly generated sequence sets. Only motifs with p-value <= 0.05 were kept. All selected motifs were used as the inputs for Promoclust. The parameters setting for Promoclust was defined as 1) minimal cluster length was 500bp, 2) minimal motifs in the cluster was two, 3) minimal conservation was 3 occurrences.

4.2.6 Generate enhancer sites graph

Graphs were generated using Perl GD.pm module. Selected motifs were represented in different colors. For each motif, the vertical bar represents the length of the motif and the horizontal bar represents the location of the motif on the sequence. The identified potential enhancer sites were highlighted in red box.

4.3 Results

4.3.1 Identify local permutation clusters (LPC) for Scr regulated genes

There are various classes of transcriptional regulatory elements, such as core promoters, proximal promoters, enhancers, silencers and insulator. Unlike core promoters and proximal promoters, which are about less than 1kb away from transcriptional start site (TSS), enhancers are often found far away from TSS. In some extreme cases, they were found 1Mb bps away from the promoter. As discussed in the introduction section, enhancers were also found in introns and downstreams. The wide range of enhancer locations makes the detection of true enhancer sites extremely difficult. Therefore, most existing algorithms only focused on a small portion of promoter to search for potential enhancer sites (Sharan, Ovcharenko et al. 2003). Although this approach decreased false positive predictions, a lot of true enhancer sites were missed too. One of the reasons that they didn't extend the search space is because their pattern discovery approach was based on alignment. There are two limitations of alignment-based approaches. First, the longer the sequences are and the further the species are apart, the harder to align them correctly. There are not only a lot of gaps between sequences, but also a lot of missing components from species to species. Second, due to TFBS turnover, even though some TFBSs are conserved across species, they couldn't be detected by alignment. We solved these two barriers but applying a non-alignment based pattern discovery method. First, to maximally retrieve all possible enhancer sites, we defined a large search space for each target gene (**Figure 4-4**). For each target gene, we selected three regions, upstream, intron and downstream. For upstream section, it was defined as the intergenic sequence between target gene's TSS and previous neighboring

gene's 3'end if the length was no longer than 35kb, or the entire intergenic sequence if the length is larger than 10kb but smaller than 35kb. If the length was less than 10kb, at least 10kb of sequence from TSS was selected. For the intron section, we selected all introns in the gene if there was any. For downstream section, it was defined similar to upstream section except the distance is between target gene's 3'end and the next gene's TSS. By applying this approach, each target gene had at least two regions that might harbor enhancers. All repeats and coding regions in these sequences were mask to increase signal to noise ratio. To detect the evolutionarily conserved regions in target genes in *D. melanogaster*, we first retrieved their corresponding orthologous sequences from 11 other fly genomes (**Table 4-2**). In the previous chapter, we shown that SPLASH algorithm (Califano 2000) enabled us to identify highly conserved patterns (SCMs) in a set of sequences without the alignment requirement. We applied it to our gene orthologous sequence set here. Multiple SPLASH parameters combination were tested and we found when minimal pattern size was set as 8bp, NT kernel density was set as 75% and orthologous conservation percentage was set as 50%, we got best coverage for all selected sequences. Top 15% or 1500bp (whichever is bigger) of most conserved SCMs were selected (please check **Materials and Methods** for details). These sequences showed highly conservation across orthologous genomes and were used as the inputs for PromoClust to identify local permutation clusters (LPCs) in each ortholog sequence set. For our positive control, fkh gene, the known enhancer was among the most conserved SCMs.

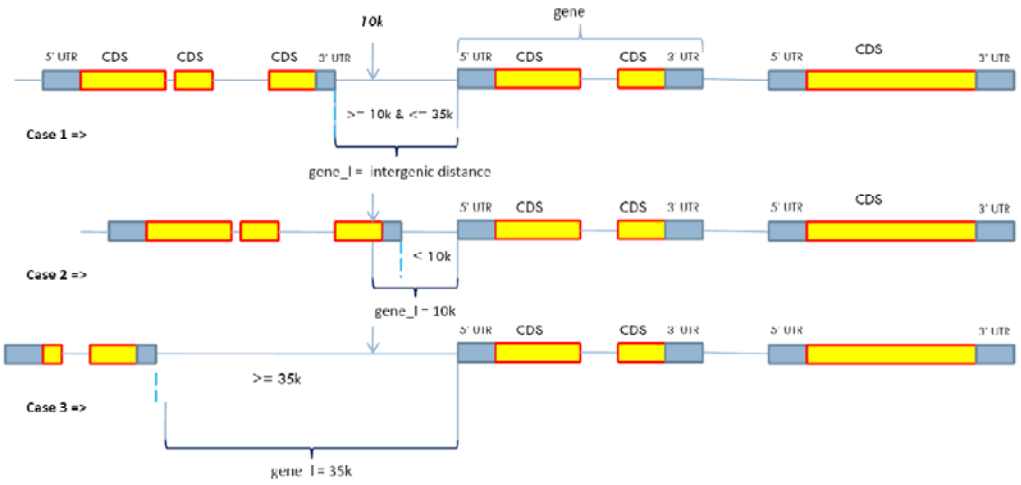


Figure 4-4. Criteria for sequences selection

Table 4-2. 12 Fly genomes

Fly Genomes
D. melanogaster
D. pseudoobscura
D. sechellia
D. ananassae
D. erecta
D. grimshawi
D. yakuba
D. mojavensis
D. persimilis
D. simulans
D. virilis
D. willistoni

For each reported LPC, it contained at least two distinct SCMs and found in at least 50% of the orthologous sequences of the given gene. Conservation score of identified LPCs were calculated based on the sum of all conserved nucleotides found in the SCMs that consists the LPCs (**Figure 4-5**). If identified LPCs had overlaps, they were combined into one piece. By applying these criteria, 44 LPCs were identified from 14 target genes' upstream, intron and downstream sequences. Because we were interested in studying the enhancer sites in *D. melanogaster*, only LPCs from *D. melanogaster's* genome were selected for the next analysis. We showed that the known Scr enhancer site on fkh gene was in one the LPCs. This also suggested that the true common enhancers for Scr target genes were harbored in these selected LPCs. Most LPCs were from upstream or downstream of the target genes, only two genes' LPCs were from intron regions. The statistical significances of LPCs were estimated according to the scoring function described in Sosinsky *et al*'s previous work (Sosinsky, Honig et al. 2007).

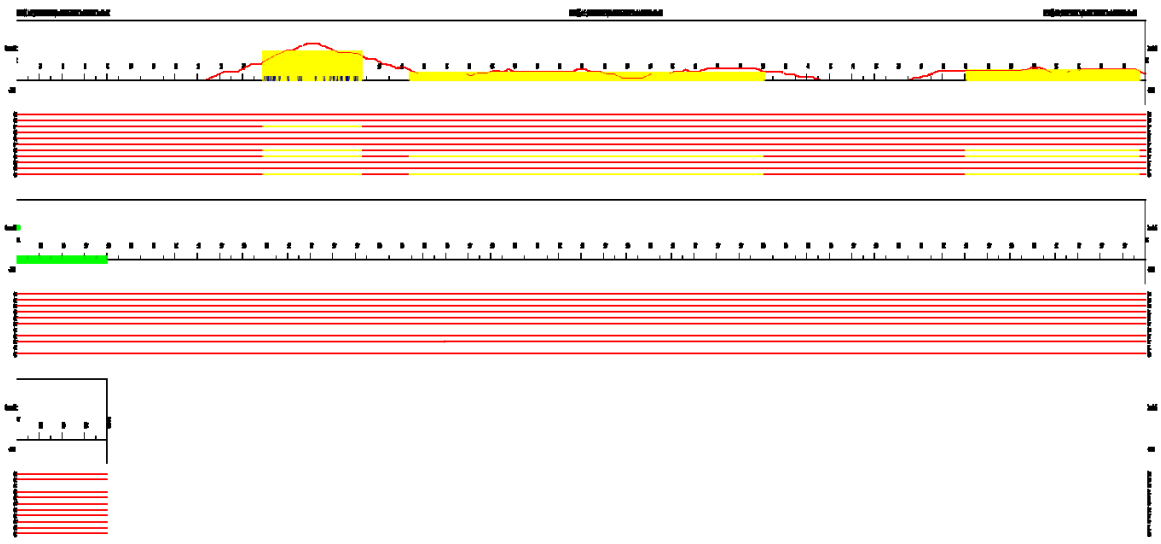


Figure 4-5. Demonstration of LPC conservation scores based sum of all conserved SCM nucleotides

4.3.2 Cluster all LPCs from target genes based on shared highly evolutionarily conserved patterns

By applying non-alignment based comparative genomics approach, we successfully identified highly evolutionarily conserved regions in each target gene. But not all of these LPCs contain the same CRMs. On the other hand, because these genes were all regulated by Scr, we assumed that at least some of LPCs shared the common enhancer sites that harbor Scr binding sites. These LPCs should thus be more similar to each other than to other conserved LPCs (**Figure 4-6**). We thus used shared SCMs as the criterion to further cluster LPCs into different subgroups. The clustering process was following. First, LPC_i was compared to all $LPC_{\neg i}$ in a 200bp window and the number of statistically significant shared SCMs were counted (please check **Materials and Methods** for details). The similarity score between two LPCs was defined as the sum of all statistically significant 200bp similarity score. 44 LPCs were clustered into several subgroups and only the statistically significant LPC subgroups (p-value ≤ 0.05, compared to random generated groups) were selected for binding sites prediction.

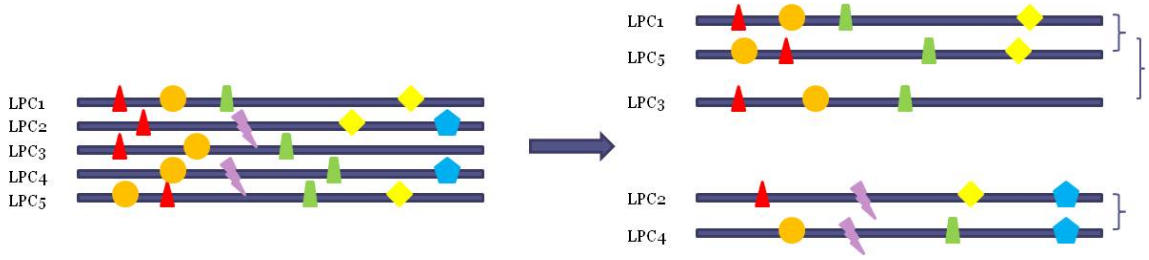


Figure 4-6. Demonstration of the hypothesis for clustering LPCs

4.3.3 Predict most enriched motifs in selected LPC group

By clustering LPCs into subgroups based there common conserved patterns, we have identified the LPCs subgroups that likely contained enhancers that harbor Scr binding site. But the since the binding motif of Scr was not available, we couldn't scan all LPCs to locate the potential binding site for Scr. To overcome this barrier, we used DME, a motif discovery method discussed in the previous chapter (Smith, Sumazin et al. 2005). We hypothesized that since these target gene were regulated by Scr, Scr's binding motif should be one the most enriched motifs discovered in these LPCs. Scr is a member of Hox family. Hox genes play important roles during embryos development in multicellular animals (McGinnis and Krumlauf 1992). It has been shown that there are 8 Hox genes in the fruit fly and 40 in mouse (**Figure 4-7**)(Chopra). All Hox proteins bind to their target genes using a domain known as the homeodomain. But the specificity of Hox proteins to their targets is very low. Nearly all Hox proteins bind to very similar DNA sequences *in vitro* (Noyes, Christensen et al. 2008). In order to specifically bind to the target genes at the right and right position, Hox proteins need an 'assistant'. In flies, the PBC factor Extradenticle (Exd) plays this role. When Exd was present, Scr showed stronger binding affinity to it target gene, fkh (Ryoo and Mann 1999; Andrew, Henderson et al. 2000) than another Hox proteins. The consensus binding pattern of Scr-Exd looked like 'TGATNNATNN'. We *de novo* predicted enriched motifs and scan for enriched known motifs. It shown that 'TNNT' pattern were identified as one the most enriched motifs by either the *de novo* prediction or scanning with known TF binding motifs (**Figure 4-8**). This suggested that binding sites for Hox transcriptional factors were enriched in the selected LPCs.

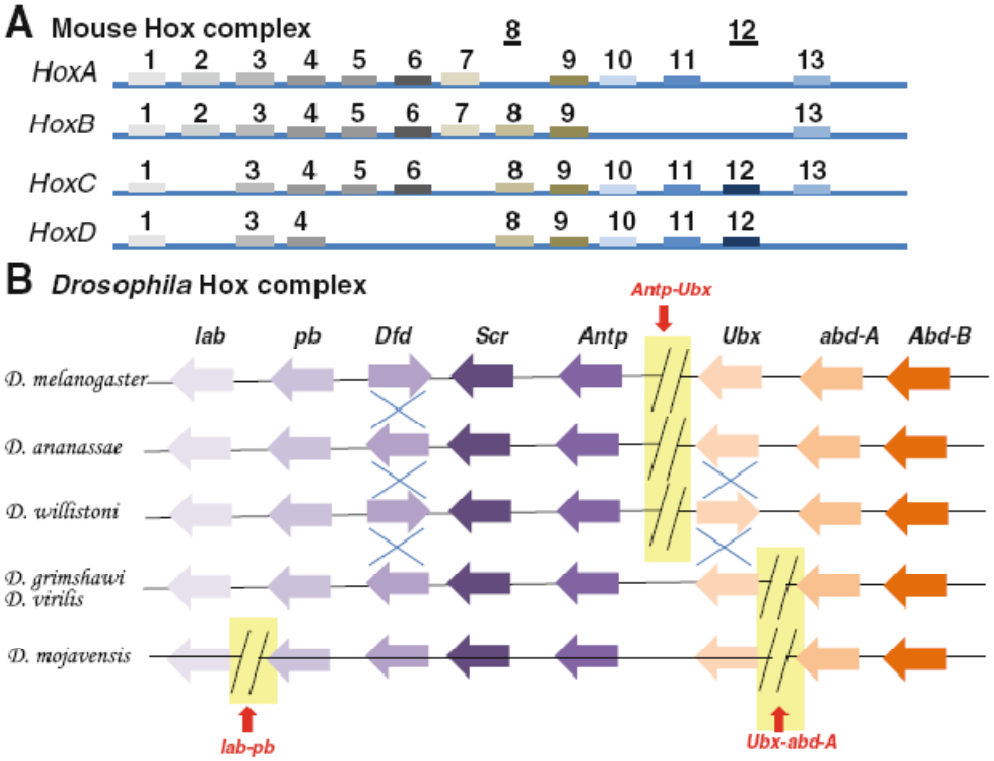


Figure 4-7. Hox complex of Mouse and Drosophila (Chopra).

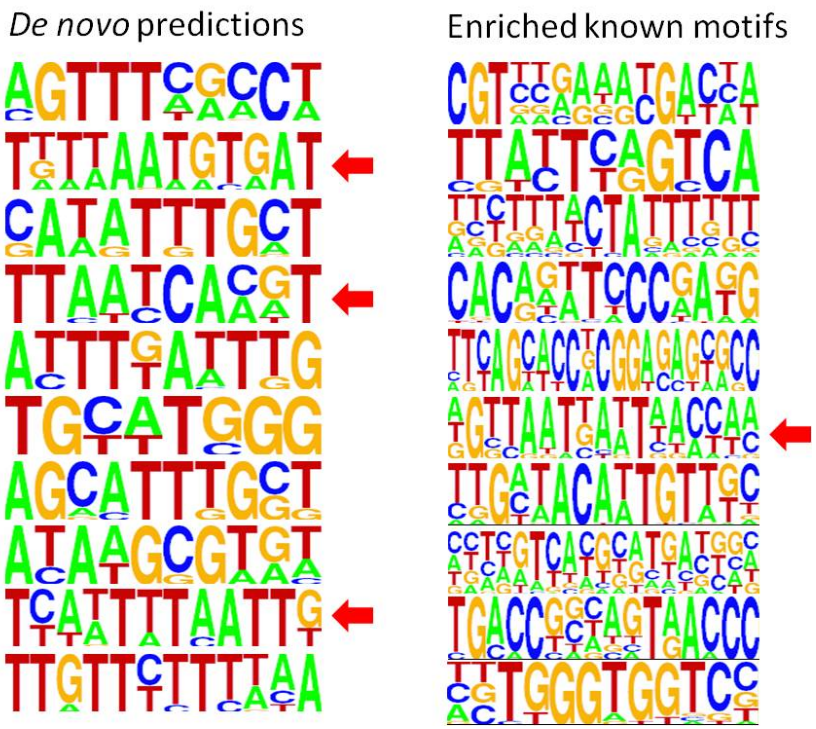


Figure 4-8. Predicted most enriched motifs in LPCs.

4.3.4 Detect enhancer site in each target gene for Scr

As described above, enhancers are usually 300 to 1000 bps long and contain 2 to 12 TF binding sites. Although enhancers might be conserved across species, the order of the TFBSs might not. In addition, the number of TFBSs in the enhancer might be different. Promoclust method enabled us to define the minimal number of conserved patterns in the cluster and the size of the cluster window (Please check **Materials and Methods** for details). We applied Promoclust again using predicted enriched motifs to look for the clusters that meet the enhancer criteria. For each predicted motifs from DME, we defined a 200bp window and required at least two motif binding sites in the window. All clusters were ranked based on the number of motifs inside and the number of

appearances in the selected LPCs. Motifs were represented with distinct colors and identified clusters were highlighted with red box as demonstrated in **Figure 4-9**.

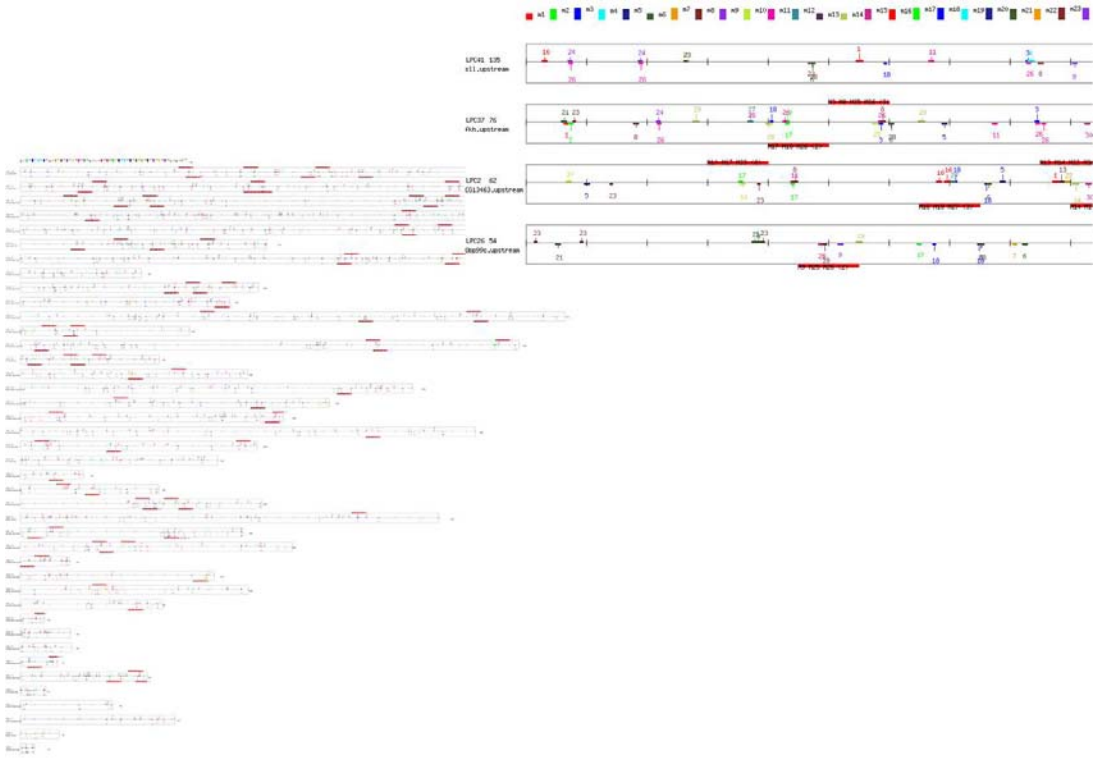


Figure 4-9. Demonstration of predicted enhancers for Scr in the target genes.

4.3.5 Validation of predicted enhancer sites

We first checked whether previous known enhancer site on fkh gene was found. Results showed that it was in one of the identified enhancers. We also generated a list of potential enhancer sites and these sites will be tested by *in vivo* gene report assay (**Appendix II**).

4.4 Discussion

In this project, we developed an enhancer discovery algorithm. This algorithm was based the previous work of Sosinsky *et al.*(Sosinsky, Honig et al. 2007). In their previous study, they showed that large regulatory elements could be identified by using genomic information alone. Using non-alignment based approach enabled them to start with a relative large initial search space for potential CRMs compared to other algorithms. Although this is an encouraging discovery, there are some limitations in their approach. First, their algorithm only allows them to study one gene at a time. All the significances were simply depended on conservations across orthologous species. But in reality, functional important sites might not conserve well across all species, especially in distal species. Second, the resolution of their approach was not very high. They could identify large pieces of regulatory elements but wouldn't allow users to narrow the search to a relatively small region which is needed for experimental validation. Third, the identified regulatory elements were not specific associated with a known TF and users needed further analysis to identify the association between predicted CRMs and TF. coEDGi solved these limitations by introducing gene co-regulation information and associating the analysis to a TF of interest. In addition, orthologous species were increased from 7 in Sosinsky *et al.*'s study to 11 in coEDGi. The resolution in coEDGi was also increased to hundred basepairs which make *in vivo* experimental validation plausible.

In this project, we tested coEDGi on Scr. The first step was to select the potential target genes that were regulated by Scr. 13 genes that were up-regulated by Scr but down-regulated by other Hox proteins were selected. This suggested that Scr binding sites might locate in these genes' enhancers. In addition, 8 of these genes were likely expressed in salivary gland, 6 out of these 8 were known to be in salivary gland and 4 of out of the 6 were known involved in sulfation pathway. Scr was known tightly linked to salivary gland formation in the *Drosophila* embryo. When Scr function was missing, salivary glands couldn't form, and if Scr was expressed everywhere, salivary glands form in new places (Andrew, Henderson et al. 2000). In order to quickly estimate coEDGi predicted enhancers, fkh gene was added to the target gene list, because a known Scr enhancer site has been previously annotated on fkh promoter region ((Ryoo and Mann 1999). We shown that coEDGi accurately discovered this known enhancer site.

coEDGi has some limitations too. First, a prior knowledge, such as TF target gene set, is required. The accuracy of the prediction largely dependent on the reliability of the input target gene set. In the future, this target can either be generated by co-regulation followed by curation or by iARACNe algorithm as described in the previous chapter. Second, although when local permutation clusters (LPCs) were identified by non-alignment based method, the conservation was still based on evolution. This would still miss some of the poorly-conserved enhancers. An alternative way is to apply a phylogenetic shadowing approach by selecting closely related species and identify additional conservation sites. Reserved patterns from both should be combined to reduce the possibility of missing some important enhancer sites.

In summary, coEDGi is a TF-centric enhancer detecting algorithm and enables users to quickly identify potential enhancer sites that harbor the binding sites of the given TF. In addition, when combined with iARACNe, users can apply genome-wide detection for all TFs with valid target gene sets.

Reference

(2008). "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." Nature **455**(7216): 1061-8.

Aach, J., W. Rindone, et al. (2000). "Systematic management and analysis of yeast gene expression data." Genome Res **10**(4): 431-45.

Adams, M. D., S. E. Celniker, et al. (2000). "The genome sequence of Drosophila melanogaster." Science **287**(5461): 2185-95.

Andrew, D. J., K. D. Henderson, et al. (2000). "Salivary gland development in Drosophila melanogaster." Mech Dev **92**(1): 5-17.

Arnone, M. I. and E. H. Davidson (1997). "The hardwiring of development: organization and function of genomic regulatory systems." Development **124**(10): 1851-64.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Asif, H. M., M. D. Rolfe, et al. "TFInfer: a tool for probabilistic inference of transcription factor activities." Bioinformatics **26**(20): 2635-6.

Bader, G. D., D. Betel, et al. (2003). "BIND: the Biomolecular Interaction Network Database." Nucleic Acids Res **31**(1): 248-50.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol **2**: 28-36.

Bailey, T. L. and C. Elkan (1995). "The value of prior knowledge in discovering motifs with MEME." Proc Int Conf Intell Syst Mol Biol **3**: 21-9.

Banerji, J., S. Rusconi, et al. (1981). "Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences." Cell **27**(2 Pt 1): 299-308.

Bansal, M. and D. di Bernardo (2007). "Inference of gene networks from temporal gene expression profiles." IET Syst Biol **1**(5): 306-12.

Barabasi, A. L. and Z. N. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nat Rev Genet **5**(2): 101-13.

Barenco, M., D. Tomescu, et al. (2006). "Ranked prediction of p53 targets using hidden variable dynamic modeling." Genome Biol **7**(3): R25.

Basso, K., A. Margolin, et al. (2005). "Reverse engineering of regulatory networks in human B cells." Nat Genet **37**(4): 382-90.

Basso, K., A. A. Margolin, et al. (2005). "Reverse engineering of regulatory networks in human B cells." Nat Genet **37**(4): 382-90.

Basso, K., M. Saito, et al. "Integrated biochemical and computational approach identifies BCL6 direct target genes controlling multiple pathways in normal germinal center B cells." Blood **115**(5): 975-84.

Beer, M. A. and S. Tavazoie (2004). "Predicting gene expression from sequence." Cell **117**(2): 185-98.

Bell, A. C., A. G. West, et al. (2001). "Insulators and boundaries: versatile regulatory elements in the eukaryotic." Science **291**(5503): 447-50.

Benos, P. V., M. L. Bulyk, et al. (2002). "Additivity in protein-DNA interactions: how good an approximation is it?" Nucleic Acids Res **30**(20): 4442-51.

Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Blackwood, E. M. and J. T. Kadonaga (1998). "Going the distance: a current view of enhancer action." Science **281**(5373): 60-3.

Blanchette, M., W. J. Kent, et al. (2004). "Aligning multiple genomic sequences with the threaded blockset aligner." Genome Res **14**(4): 708-15.

Blanchette, M. and S. Sinha (2001). "Separating real motifs from their artifacts." Bioinformatics **17 Suppl 1**: S30-8.

Blanchette, M. and M. Tompa (2002). "Discovery of regulatory elements by a computational method for phylogenetic footprinting." Genome Res **12**(5): 739-48.

Blencowe, B., S. Brenner, et al. (2009). "Post-transcriptional gene regulation: RNA-protein interactions, RNA processing, mRNA stability and localization." Pac Symp Biocomput: 545-8.

Boffelli, D., J. McAuliffe, et al. (2003). "Phylogenetic shadowing of primate sequences to find functional regions of the human genome." Science **299**(5611): 1391-4.

Boulesteix, A. L. and K. Strimmer (2005). "Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach." Theor Biol Med Model **2**: 23.

Boyadjiev, S. A. and E. W. Jabs (2000). "Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders." Clin Genet **57**(4): 253-66.

Brand, A. H., L. Breeden, et al. (1985). "Characterization of a "silencer" in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer." Cell **41**(1): 41-8.

Brazma, A., I. Jonassen, et al. (1996). "Discovering patterns and subfamilies in biosequences." Proc Int Conf Intell Syst Mol Biol **4**: 34-43.

Brown, C. T. (2008). "Computational approaches to finding and analyzing cis-regulatory elements." Methods Cell Biol **87**: 337-65.

Bulyk, M. L., P. L. Johnson, et al. (2002). "Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors." Nucleic Acids Res **30**(5): 1255-61.

Bussemaker, H. J., H. Li, et al. (2001). "Regulatory element detection using correlation with expression." Nat Genet **27**(2): 167-71.

Butte, A. J. and I. S. Kohane (2000). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements." Pac Symp Biocomput: 418-29.

Butte, A. J., P. Tamayo, et al. (2000). "Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks." Proc Natl Acad Sci U S A **97**(22): 12182-6.

Califano, A. (2000). "SPLASH: structural pattern localization analysis by sequential histograms." Bioinformatics **16**(4): 341-57.

Carro, M., W. Lim, et al. (2009). "The transcriptional network for mesenchymal transformation of brain tumours." Nature.

Carro, M. S., W. K. Lim, et al. "The transcriptional network for mesenchymal transformation of brain tumours." Nature **463**(7279): 318-25.

Cheadle, C., Y. S. Cho-Chung, et al. (2003). "Application of z-score transformation to Affymetrix data." Appl Bioinformatics **2**(4): 209-17.

Chen, K. and N. Rajewsky (2007). "The evolution of gene regulation by transcription factors and microRNAs." Nat Rev Genet **8**(2): 93-103.

Chopra, V. S. "Chromosomal organization at the level of gene complexes." Cell Mol Life Sci.

Conlon, E. M., X. S. Liu, et al. (2003). "Integrating regulatory motif discovery and genome-wide expression analysis." Proc Natl Acad Sci U S A **100**(6): 3339-44.

Cooper, G. M., E. A. Stone, et al. (2005). "Distribution and intensity of constraint in mammalian genomic sequence." Genome Res **15**(7): 901-13.

D'Haeseleer, P., S. Liang, et al. (2000). "Genetic network inference: from co-expression clustering to reverse engineering." Bioinformatics **16**(8): 707-26.

D'Haeseleer, P., X. Wen, et al. (1999). "Linear modeling of mRNA expression levels during CNS development and injury." Pac Symp Biocomput: 41-52.

Day, D. A. and M. F. Tuite (1998). "Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview." J Endocrinol **157**(3): 361-71.

di Bernardo, D., M. J. Thompson, et al. (2005). "Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks." Nat Biotechnol **23**(3): 377-83.

Elnitski, L., V. X. Jin, et al. (2006). "Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques." Genome Res **16**(12): 1455-64.

Elnitski, L., D. King, et al. (2006). "Computational prediction of cis-regulatory modules from multispecies alignments using Galaxy, Table Browser, and GALA." Methods Mol Biol **338**: 91-103.

Erives, A. and M. Levine (2004). "Coordinate enhancers share common organizational features in the Drosophila genome." Proc Natl Acad Sci U S A **101**(11): 3851-6.

Farnham, P. J. (2009). "Insights from genomic profiling of transcription factors." Nat Rev Genet **10**(9): 605-16.

Filipowicz, W., S. N. Bhattacharyya, et al. (2008). "Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?" Nat Rev Genet **9**(2): 102-14.

Friedman, N., M. Linial, et al. (2000). "Using Bayesian networks to analyze expression data." J Comput Biol **7**(3-4): 601-20.

Fu, N., I. Drinnenberg, et al. (2007). "Comparison of protein and mRNA expression evolution in humans and chimpanzees." PLoS One **2**(2): e216.

Furney, S. J., D. G. Higgins, et al. (2006). "Structural and functional properties of genes involved in human cancer." BMC Genomics **7**: 3.

Gao, F., B. C. Foat, et al. (2004). "Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data." BMC Bioinformatics **5**: 31.

Gardner, T. S., D. di Bernardo, et al. (2003). "Inferring genetic networks and identifying compound mode of action via expression profiling." Science **301**(5629): 102-5.

Hart, R. K., A. K. Royyuru, et al. (2000). "Systematic and fully automated identification of protein sequence patterns." J Comput Biol **7**(3-4): 585-600.

Hartemink, A. J., D. K. Gifford, et al. (2002). "Combining location and expression data for principled discovery of genetic regulatory network models." Pac Symp Biocomput: 437-49.

Hughes, J. D., P. W. Estep, et al. (2000). "Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae." J Mol Biol **296**(5): 1205-14.

Hunter, S., R. Apweiler, et al. (2009). "InterPro: the integrative protein signature database." Nucleic Acids Res **37**(Database issue): D211-5.

Imoto, S., T. Higuchi, et al. (2003). "Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks." Proc IEEE Comput Soc Bioinform Conf **2**: 104-13.

Kawamata, N., T. Miki, et al. (1994). "Recognition DNA sequence of a novel putative transcription factor, BCL6." Biochem Biophys Res Commun **204**(1): 366-74.

Keles, S., M. van der Laan, et al. (2002). "Identification of regulatory elements using a feature selection method." Bioinformatics **18**(9): 1167-75.

Kim, T. H., Z. K. Abdullaev, et al. (2007). "Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome." Cell **128**(6): 1231-45.

Kouskouti, A., E. Scheer, et al. (2004). "Gene-specific modulation of TAF10 function by SET9-mediated methylation." Mol Cell **14**(2): 175-82.

Lachmann, A., H. Xu, et al. "ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments." Bioinformatics **26**(19): 2438-44.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lawrence, C. E., S. F. Altschul, et al. (1993). "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment." Science **262**(5131): 208-14.

Lee, J. Y., J. Colinas, et al. (2006). "Transcriptional and posttranscriptional regulation of transcription factor expression in Arabidopsis roots." Proc Natl Acad Sci U S A **103**(15): 6055-60.

Lefebvre, C., P. Rajbhandari, et al. "A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers." Mol Syst Biol **6**: 377.

Liao, J. C., R. Boscolo, et al. (2003). "Network component analysis: reconstruction of regulatory signals in biological systems." Proc Natl Acad Sci U S A **100**(26): 15522-7.

Lim, W. K., E. Lyashenko, et al. (2009). "Master regulators used as breast cancer metastasis classifier." Pac Symp Biocomput: 504-15.

Man, T. K. and G. D. Stormo (2001). "Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay." Nucleic Acids Res **29**(12): 2471-8.

Margolin, A. A., I. Nemenman, et al. (2006). "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context." BMC Bioinformatics **7 Suppl 1**: S7.

Margolin, A. A., K. Wang, et al. (2006). "Reverse engineering cellular networks." Nat Protoc **1**(2): 662-71.

Markstein, M. and M. Levine (2002). "Decoding cis-regulatory DNAs in the Drosophila genome." Curr Opin Genet Dev **12**(5): 601-6.

Maston, G. A., S. K. Evans, et al. (2006). "Transcriptional regulatory elements in the human genome." Annu Rev Genomics Hum Genet **7**: 29-59.

Matys, V., E. Fricke, et al. (2003). "TRANSFAC: transcriptional regulation, from patterns to profiles." Nucleic Acids Res **31**(1): 374-8.

McGinnis, W. and R. Krumlauf (1992). "Homeobox genes and axial patterning." Cell **68**(2): 283-302.

Moses, A. M., D. Y. Chiang, et al. (2004). "MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model." Genome Biol **5**(12): R98.

Noyes, M. B., R. G. Christensen, et al. (2008). "Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites." Cell **133**(7): 1277-89.

Opper, M. and G. Sanguinetti "Learning combinatorial transcriptional dynamics from gene expression data." Bioinformatics **26**(13): 1623-9.

Perrin, B. E., L. Ralaivola, et al. (2003). "Gene networks inference using dynamic Bayesian networks." Bioinformatics **19 Suppl 2**: ii138-48.

Phan, R. T., M. Saito, et al. (2005). "BCL6 interacts with the transcription factor Miz-1 to suppress the cyclin-dependent kinase inhibitor p21 and cell cycle arrest in germinal center B cells." Nat Immunol **6**(10): 1054-60.

Polo, J. M., P. Juszczynski, et al. (2007). "Transcriptional signature with differential expression of BCL6 target genes accurately identifies BCL6-dependent diffuse large B cell lymphomas." Proc Natl Acad Sci U S A **104**(9): 3207-12.

Ryoo, H. D. and R. S. Mann (1999). "The control of trunk Hox specificity and activity by Extradenticle." Genes Dev **13**(13): 1704-16.

Sabatti, C. and G. M. James (2006). "Bayesian sparse hidden components analysis for transcription regulation networks." Bioinformatics **22**(6): 739-46.

Sadler, J. R., M. S. Waterman, et al. (1983). "Regulatory pattern identification in nucleic acid sequences." Nucleic Acids Res **11**(7): 2221-31.

Sandelin, A., W. Alkema, et al. (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic Acids Res **32**(Database issue): D91-4.

Sanguinetti, G., N. D. Lawrence, et al. (2006). "Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities." Bioinformatics **22**(22): 2775-81.

Schones, D. E., P. Sumazin, et al. (2005). "Similarity of position frequency matrices for transcription factor binding sites." Bioinformatics **21**(3): 307-13.

Schroeder, M. D., M. Pearce, et al. (2004). "Transcriptional control in the segmentation gene network of Drosophila." PLoS Biol **2**(9): E271.

Schultheiss, S. J., W. Busch, et al. (2009). "KIRMES: kernel-based identification of regulatory modules in euchromatic sequences." Bioinformatics **25**(16): 2126-33.

Sharan, R., I. Ovcharenko, et al. (2003). "CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments." Bioinformatics **19 Suppl 1**: i283-91.

Siepel, A., G. Bejerano, et al. (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." Genome Res **15**(8): 1034-50.

Sinha, S. and M. Tompa (2003). "YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation." Nucleic Acids Res **31**(13): 3586-8.

Smith, A. D., P. Sumazin, et al. (2005). "Mining ChIP-chip data for transcription factor and cofactor binding sites." Bioinformatics **21 Suppl 1**: i403-12.

Smith, A. D., P. Sumazin, et al. (2005). "Identifying tissue-selective transcription factor binding sites in vertebrate promoters." Proc Natl Acad Sci U S A **102**(5): 1560-5.

Smith, A. D., P. Sumazin, et al. (2007). "Tissue-specific regulatory elements in mammalian promoters." Mol Syst Biol **3**: 73.

Sosinsky, A., B. Honig, et al. (2007). "Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting." Proc Natl Acad Sci U S A **104**(15): 6305-10.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

Stormo, G. D. and G. W. Hartzell, 3rd (1989). "Identifying protein-binding sites from unaligned DNA fragments." Proc Natl Acad Sci U S A **86**(4): 1183-7.

Subramanian, A., P. Tamayo, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-50.

Sumazin, P., G. Chen, et al. (2005). "DWE: discriminating word enumerator." Bioinformatics **21**(1): 31-8.

Tagle, D. A., B. F. Koop, et al. (1988). "Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints." J Mol Biol **203**(2): 439-55.

Tompa, M., N. Li, et al. (2005). "Assessing computational tools for the discovery of transcription factor binding sites." Nat Biotechnol **23**(1): 137-44.

Vaquerizas, J. M., S. K. Kummerfeld, et al. (2009). "A census of human transcription factors: function, expression and evolution." Nat Rev Genet **10**(4): 252-63.

Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Wang, K., M. Saito, et al. (2009). "Genome-wide identification of post-translational modulators of transcription factor activity in human B cells." Nat Biotechnol **27**(9): 829-39.

Wang, T. and G. D. Stormo (2003). "Combining phylogenetic data with co-regulated genes to identify regulatory motifs." Bioinformatics **19**(18): 2369-80.

Wang, W., J. M. Cherry, et al. (2002). "A systematic approach to reconstructing transcription networks in Saccharomycescerevisiae." Proc Natl Acad Sci U S A **99**(26): 16893-8.

Ward, L. D. and H. J. Bussemaker (2008). "Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences." Bioinformatics **24**(13): i165-71.

Wasserman, W. W., M. Palumbo, et al. (2000). "Human-mouse genome comparisons to locate regulatory sites." Nat Genet **26**(2): 225-8.

Waterston, R. and J. Sulston (1995). "The genome of Caenorhabditis elegans." Proc Natl Acad Sci U S A **92**(24): 10836-40.

Xie, D., J. Cai, et al. (2008). "Cross-species de novo identification of cis-regulatory modules with GibbsModule: Application to gene regulation in embryonic stem cells." Genome Res.

Xie, X., T. S. Mikkelsen, et al. (2007). "Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites." Proc Natl Acad Sci U S A **104**(17): 7145-50.

Yoseph, B., B. Gill, et al. (2001). A Simple Hyper-Geometric Approach for Discovering Putative Transcription Factor Binding Sites. Proceedings of the First International Workshop on Algorithms in Bioinformatics, Springer-Verlag.

Youn, A., D. J. Reiss, et al. "Learning transcriptional networks from the integration of ChIP-chip and expression data in a non-parametric model." Bioinformatics **26**(15): 1879-86.

Yu, J., V. A. Smith, et al. (2004). "Advances to Bayesian network inference for generating causal networks from observational biological data." Bioinformatics **20**(18): 3594-603.

Zahnow, C. A. (2002). "CCAAT/enhancer binding proteins in normal mammary development and breast cancer." Breast Cancer Res **4**(3): 113-21.

Zeller, K. I., A. G. Jegga, et al. (2003). "An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets." Genome Biol **4**(10): R69.

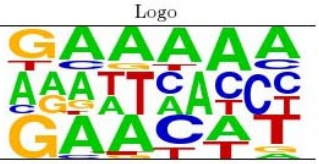
Zhang, Y., T. Liu, et al. (2008). "Model-based analysis of ChIP-Seq (MACS)." Genome Biol **9**(9): R137.

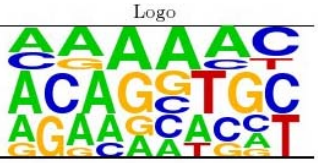
Zhou, X., P. Sumazin, et al. "A systems biology approach to transcription factor binding site prediction." PLoS One **5**(3): e9878.

Zhu, Z., Y. Pilpel, et al. (2002). "Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm." J Mol Biol **318**(1): 71-81.

Appendix I

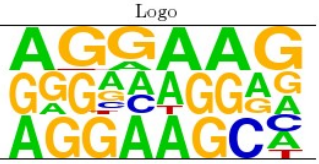
OmniMiner *de novo* predicted TFBSs for 20 selected TFs

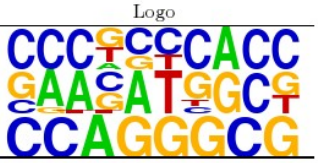
TF	Source	Err	Sens	Spec	*p*-val	Logo
BCL6	Cons	0.29	0.73	0.68	0.00	
BCL6	Cons	0.30	0.80	0.60	0.00	
BCL6	Cons	0.30	0.80	0.59	0.00	

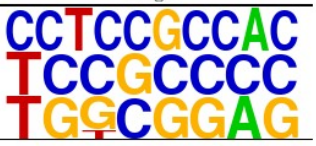
TF	Source	Err	Sens	Spec	*p*-val	Logo
NFATC1	Cons	0.27	0.78	0.69	0.00	
NFATC1	Cons	0.27	0.75	0.71	0.00	
NFATC1	Cons	0.27	0.72	0.73	0.00	

TF	Source	Err	Sens	Spec	*p*-val	Logo
NME2	Cons	0.28	0.72	0.72	0.00	
NME2	Cons	0.30	0.73	0.68	0.00	

TF	Source	Err	Sens	Spec	*p*-val	Logo
RB1	Cons	0.26	0.92	0.57	0.00	
RB1	Cons	0.28	0.70	0.73	0.00	
RB1	Cons	0.28	0.73	0.70	0.00	

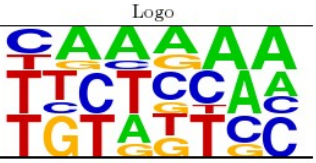
TF	Source	Err	Sens	Spec	*p*-val	Logo
VAV1	Cons	0.24	0.81	0.71	0.00	
VAV1	Cons	0.24	0.76	0.76	0.00	
VAV1	Cons	0.25	0.78	0.73	0.00	

TF	Source	Err	Sens	Spec	*p*-val	Logo
NFE2L2	Cons	0.28	0.91	0.54	0.00	
NFE2L2	Cons	0.28	0.70	0.74	0.00	
NFE2L2	Cons	0.29	0.97	0.45	0.00	

TF	Source	Err	Sens	Spec	*p*-val	Logo
MECP2	Cons	0.23	0.85	0.68	0.00	
MECP2	Cons	0.25	0.96	0.55	0.00	
MECP2	Cons	0.25	0.70	0.80	0.00	

TF	Source	Err	Sens	Spec	*p*-val	Logo
FLI1	Orig	0.28	0.69	0.75	0.00	
FLI1	Orig	0.30	0.66	0.75	0.00	
FLI1	Orig	0.30	0.77	0.64	0.00	

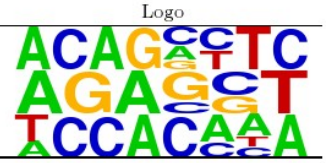
TF	Source	Err	Sens	Spec	*p*-val	Logo
HOXD13	Cons	0.32	0.73	0.63	0.00	
HOXD13	Cons	0.32	0.73	0.63	0.00	
HOXD13	Cons	0.32	0.60	0.75	0.00	

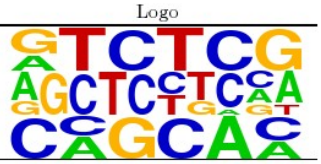
TF	Source	Err	Sens	Spec	*p*-val	Logo
HIF1A	Cons	0.27	0.71	0.76	0.00	
HIF1A	Cons	0.30	0.83	0.57	0.00	
HIF1A	Cons	0.30	0.80	0.59	0.00	

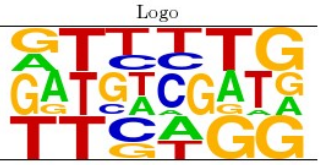
TF	Source	Err	Sens	Spec	*p*-val	Logo
ID1	Cons	0.31	0.83	0.54	0.00	
ID1	Cons	0.32	0.68	0.69	0.00	
ID1	Orig	0.33	0.58	0.77	0.00	

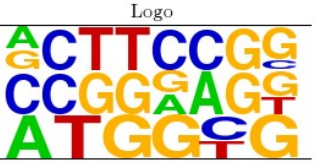
TF	Source	Err	Sens	Spec	*p*-val	Logo
MEF2C	Orig	0.29	0.77	0.64	0.00	
MEF2C	Cons	0.31	0.77	0.61	0.00	
MEF2C	Cons	0.32	0.69	0.67	0.00	

TF	Source	Err	Sens	Spec	*p*-val	Logo
PAX7	Cons	0.32	0.71	0.65	0.00	
PAX7	Cons	0.33	0.72	0.62	0.00	
PAX7	Orig	0.33	0.75	0.60	0.05	

TF	Source	Err	Sens	Spec	*p*-val	Logo
TBX1	Cons	0.25	0.74	0.76	0.00	
TBX1	Cons	0.25	0.77	0.72	0.00	
TBX1	Cons	0.27	0.74	0.72	0.00	

TF	Source	Err	Sens	Spec	*p*-val	Logo
MLL	Cons	0.32	0.76	0.60	0.00	
MLL	Cons	0.32	0.75	0.61	0.00	
MLL	Cons	0.33	0.88	0.47	0.00	

TF	Source	Err	Sens	Spec	*p*-val	Logo
APC	Cons	0.26	0.78	0.71	0.00	
APC	Cons	0.27	0.74	0.71	0.01	
APC	Cons	0.28	0.63	0.81	0.02	

TF	Source	Err	Sens	Spec	*p*-val	Logo
CEBPZ	Cons	0.27	0.72	0.75	0.00	
CEBPZ	Cons	0.27	0.84	0.61	0.00	
CEBPZ	Cons	0.28	0.73	0.72	0.00	

TF	Source	Err	Sens	Spec	*p*-val	Logo
EP300	Cons	0.26	0.84	0.65	0.00	
EP300	Cons	0.26	0.80	0.68	0.05	

TF	Source	Err	Sens	Spec	*p*-val	Logo
MSC	Cons	0.32	0.70	0.67	0.00	
MSC	Cons	0.32	0.67	0.68	0.00	
MSC	Cons	0.33	0.68	0.67	0.00	

TF	Source	Err	Sens	Spec	*p*-val	Logo
PITX1	Cons	0.25	0.83	0.67	0.00	
PITX1	Cons	0.25	0.86	0.64	0.00	
PITX1	Cons	0.25	0.86	0.63	0.00	

Appendix II

coEDGi predicted enhancer sites for Scr

NumOfOcc	Pval	Cluster
8	1.01E-29	M3 M22 M26
5	5.40E-19	M23 M26 M30
5	2.78E-19	M10 M17 M23
5	2.60E-18	M2 M3 M26
5	1.97E-18	M10 M23 M26
5	1.27E-18	M18 M20 M23
4	4.66E-15	M3 M11 M26
4	4.06E-15	M22 M23 M26
4	4.05E-16	M19 M23 M25
4	3.83E-15	M2 M10 M23
4	1.15E-15	M5 M15 M23
3	9.78E-12	M10 M11 M23
3	7.34E-12	M5 M10 M23
3	5.88E-12	M3 M21 M26
3	5.88E-12	M15 M23 M27
3	5.72E-12	M6 M18 M23
3	5.19E-12	M15 M23 M29
3	4.35E-12	M9 M26 M30
3	4.12E-12	M9 M23 M29
3	3.89E-12	M10 M11 M17
3	3.80E-12	M19 M23 M30
3	3.44E-12	M4 M11 M13
3	3.38E-12	M9 M23 M30
3	3.10E-12	M8 M16 M18
3	2.68E-11	M3 M7 M26
3	2.62E-12	M8 M23 M29
3	2.38E-11	M23 M24 M26
3	2.18E-12	M6 M9 M23
3	2.05E-12	M12 M27 M29
3	2.02E-12	M7 M17 M25
3	1.86E-12	M5 M8 M10
3	1.61E-12	M8 M23 M25
3	1.51E-11	M3 M19 M26
3	1.25E-12	M5 M8 M17
3	1.04E-11	M2 M17 M23
3	1.03E-11	M10 M20 M23
2	8.51E-09	M6 M12 M30
2	8.51E-09	M6 M11 M30
2	8.26E-11	M2 M19 M23 M25
2	7.71E-08	M11 M23 M26
2	7.55E-11	M8 M23 M25 M26
2	7.33E-08	M3 M24 M26
2	6.69E-11	M3 M8 M25 M26
2	6.56E-08	M18 M22 M26
2	6.56E-08	M17 M18 M26
2	5.90E-11	M9 M26 M29 M30

2	5.87E-08	M13 M23 M26
2	5.76E-08	M3 M12 M23
2	5.76E-08	M2 M11 M23
2	5.08E-08	M18 M26 M30
2	4.71E-08	M4 M11 M18
2	4.30E-08	M16 M24 M26
2	4.29E-08	M18 M23 M30
2	4.25E-08	M7 M13 M23
2	4.10E-08	M11 M15 M23
2	4.00E-08	M12 M26 M27
2	3.82E-08	M20 M26 M29
2	3.78E-11	M5 M8 M10 M17
2	3.59E-08	M4 M13 M18
2	3.52E-08	M5 M12 M23
2	3.39E-08	M7 M11 M28
2	3.26E-08	M3 M6 M26
2	3.24E-08	M7 M10 M22
2	3.21E-08	M3 M10 M27
2	3.08E-08	M8 M11 M26
2	2.99E-08	M14 M18 M27
2	2.96E-08	M3 M10 M29
2	2.87E-08	M5 M18 M27
2	2.87E-08	M18 M22 M27
2	2.79E-08	M8 M23 M24
2	2.78E-08	M19 M23 M29
2	2.76E-08	M16 M18 M27
2	2.75E-08	M2 M21 M23
2	2.61E-10	M3 M22 M23 M26
2	2.38E-08	M4 M15 M17
2	2.37E-08	M2 M5 M13
2	2.20E-08	M9 M25 M26
2	2.20E-08	M17 M25 M26
2	2.15E-08	M7 M8 M15
2	2.15E-08	M8 M17 M23
2	2.13E-08	M4 M14 M22
2	2.03E-10	M4 M18 M23 M28
2	1.91E-08	M6 M26 M27
2	1.90E-08	M2 M5 M8
2	1.80E-08	M6 M14 M18
2	1.75E-08	M9 M11 M13
2	1.65E-10	M3 M15 M22 M26
2	1.59E-08	M11 M19 M30
2	1.41E-08	M8 M11 M17
2	1.36E-10	M3 M22 M26 M27
2	1.34E-08	M18 M21 M30
2	1.33E-08	M17 M27 M29
2	1.30E-08	M8 M12 M13

2	1.24E-08	M6 M8 M23
2	1.03E-08	M8 M9 M29
2	1.01E-07	M7 M23 M26
1	9.39E-06	M5 M12 M17 M23
1	8.22E-06	M2 M3 M5 M25
1	8.10E-06	M3 M7 M17 M25
1	8.07E-06	M8 M11 M17 M23
1	6.01E-06	M11 M15 M21 M28
1	5.47E-06	M8 M12 M27 M29
1	5.41E-06	M13 M14 M22 M30
1	4.39E-06	M16 M21 M22 M29
1	3.88E-07	M6 M12 M18 M24 M29
1	2.03E-09	M4 M11 M13 M15 M18 M23 M28
1	1.47E-09	M3 M8 M14 M15 M19 M22 M26
1	1.11E-05	M5 M18 M23 M29
1	0.000270763	M3 M4 M26
1	0.000187483	M7 M8 M18
1	0.000162637	M18 M22 M29
1	0.000149903	M6 M10 M18
1	0.000143386	M10 M24 M30
1	0.000138214	M19 M22 M28
1	0.000137275	M19 M20 M29
1	0.000132066	M7 M11 M21
1	0.000131494	M5 M6 M18
1	0.000122768	M13 M14 M22
1	0.000121813	M21 M23 M29
1	0.000117039	M14 M19 M30
1	0.000101354	M5 M8 M29
1	0.00010085	M15 M16 M21

Coordinates

3R:24403443-24403519;3R:26315676-26315776;3R:5096701-5096801;3L:19841038-19841088;3R:19495609-1949
3R:13439923-13440023;3R:13424905-13425005;3R:26305827-26305927;3R:24403343-24403443;3R:25514560-25
3R:25514510-25514610;3R:25508004-25508104;3R:19500754-19500854;3L:15047266-15047365;3L:15041699-15(
3R:26305877-26305977;3R:24403443-24403519;3R:5096551-5096651;3R:26326009-26326109;3R:13427905-1342
3R:25514510-25514610;3R:25508004-25508104;3R:24419694-24419744;3R:19499849-19499899;3L:15044249-15(
3R:26310722-26310822;3L:15040707-15040801;3R:24417220-24417320;3R:5096351-5096451;3R:19500754-1950(
3L:13958918-13959018;3L:19840988-19841088;3R:19498049-19498149;3R:24415020-24415120
3L:15047331-15047431;3R:24418744-24418844;3R:19495609-19495709;3R:13424905-13424955
3R:19499799-19499899;3L:13971779-13971879;3L:15046431-15046531;3R:24411268-24411318
3R:19500754-19500854;3R:13425205-13425305;3R:26325959-26326059;3L:15044249-15044299
3L:15048965-15049065;3L:19863230-19863330;3L:15040657-15040757;3R:13425505-13425555
3R:25517560-25517610;3L:15047266-15047365;3L:15041699-15041799
3R:13425155-13425255;3L:15041399-15041499;3L:19863280-19863330
3R:26305877-26305977;3L:13958918-13959018;3L:15042599-15042699
3L:15047331-15047431;3L:15049015-15049115;3L:15055030-15055130
3R:19500704-19500804;3R:26322131-26322231;3R:5096351-5096401
3R:24403343-24403443;3L:15040707-15040757;3L:15055080-15055130
3R:24403343-24403443;3L:19863330-19863430;3R:24411418-24411518
3R:24403343-24403443;3L:15056837-15056937;3R:19497449-19497549
3L:15047266-15047365;3L:15041699-15041799;3R:24415820-24415870
3L:15040357-15040457;3L:15051352-15051452;3R:13424905-13425005
3L:22677888-22677988;3L:15053252-15053352;3L:15047281-15047365
3L:19863280-19863380;3R:24403343-24403443;3R:5096251-5096351
3L:15040457-15040551;3R:19501554-19501654;3L:15052452-15052552
3R:5096701-5096801;3R:24410918-24411018;3L:15053652-15053752
3R:24403343-24403443;3R:19498349-19498449;3R:25515160-25515210
3R:24419694-24419744;3R:26305827-26305927;3R:24417620-24417670
3R:24403343-24403443;3L:15056837-15056937;3L:13973429-13973529
3L:15044299-15044399;3L:13968042-13968142;3R:13424755-13424805
3R:24410968-24411068;3R:26305977-26306077;3R:19495259-19495309
3R:25518160-25518260;3L:22684419-22684519;3R:13427605-13427705
3R:24403343-24403443;3R:26310622-26310722;3L:15040857-15040901
3R:13424905-13425005;3R:26315726-26315793;3R:13438573-13438623
3R:13427605-13427705;3L:22684369-22684469;3R:25518210-25518260
3R:26308872-26308972;3R:13425905-13426005;3R:19500754-19500804
3R:24415920-24415970;3R:19500754-19500854;3R:26325959-26326059
3R:13439023-13439123;3R:24403343-24403443
3L:22685991-22686091;3R:26320131-26320231
3R:24411268-24411368;3L:15046431-15046531
3L:15047331-15047431;3L:13966029-13966129
3R:24403343-24403443;3L:15040857-15040901
3R:26305827-26305927;3L:15052102-15052202
3R:26326059-26326109;3R:24414820-24414870
3L:15039107-15039207;3R:26315676-26315776
3R:24417970-24418070;3R:24414670-24414720
3R:24403343-24403443;3L:19863330-19863430

3R:24418694-24418794;3R:24417620-24417720
3R:19501754-19501804;3L:15056587-15056637
3R:19495659-19495759;3L:15053102-15053202
3R:26308772-26308872;3L:19839038-19839138
3L:22683669-22683769;3L:15047281-15047365
3R:13423155-13423255;3L:19863130-19863180
3R:26310722-26310822;3R:19497649-19497749
3R:24417620-24417720;3R:19501854-19501904
3R:25515960-25516060;3L:15047331-15047365
3R:5096001-5096101;3L:15044249-15044249
3R:13439973-13440065;3L:15056887-15056987
3R:25518160-25518260;3R:13427605-13427705
3L:22677838-22677938;3L:15047281-15047331
3R:13425155-13425205;3R:5096951-5097001
3L:15043949-15044049;3L:15056637-15056737
3L:13958968-13959068;3R:24414820-24414920
3R:5096701-5096801;3R:19501954-19502054
3L:19838738-19838838;3R:5096701-5096801
3R:25518210-25518310;3L:13972879-13972979
3R:13426305-13426405;3L:13968042-13968142
3R:25518110-25518210;3R:5096701-5096801
3L:13958718-13958818;3L:15042049-15042149
3R:26311122-26311172;3R:13427305-13427405
3L:15056487-15056587;3L:22682969-22683069
3R:19499799-19499899;3R:13425605-13425705
3L:15040457-15040551;3L:15042028-15042078
3R:26321131-26321231;3R:24414320-24414370
3R:19495609-19495709;3R:13424905-13424955
3L:15047266-15047365;3L:22683619-22683719
3R:19499549-19499649;3R:13439423-13439523
3R:24403343-24403443;3R:25513760-25513860
3R:25507954-25508054;3R:26305977-26306077
3R:25518260-25518346;3R:25517060-25517160
3L:13973679-13973779;3L:15041728-15041799
3L:19862080-19862180;3L:15047531-15047631
3L:15044055-15044149;3L:15047281-15047365
3R:5096651-5096751;3R:19495359-19495459
3R:13424305-13424355;3L:15048965-15049015
3L:13968042-13968142;3R:19512214-19512264
3L:15053202-15053302;3R:26325109-26325159
3R:19495559-19495659;3R:13438573-13438623
3L:15040407-15040457;3R:26311972-26312022
3R:25518210-25518310;3L:15041778-15041799
3L:19841038-19841138;3R:5096701-5096801
3L:15039307-15039407;3L:19839038-19839138
3R:24414620-24414720;3L:22678038-22678138
3R:19499599-19499699;3L:15040457-15040551

3L:15042378-15042478;3R:24403343-24403443
3R:24403343-24403443;3R:19497499-19497599
3R:24417620-24417720;3L:22685741-22685841
3R:5096951-5097001
3L:15046481-15046531
3R:19495259-19495309
3L:15041728-15041799
3L:13958918-13958968
3R:13424755-13424805
3L:15042228-15042299
3L:22686991-22687041
3L:13968042-13968092
3L:15047281-15047365
3R:13438573-13438623
3L:15040657-15040751
3R:13423955-13424005
3L:15046331-15046381
3L:22681919-22681969
3R:26321931-26321981
3L:13967942-13967992
3L:15043955-15044049
3L:13968192-13968242
3L:13968242-13968292
3L:15042099-15042149
3L:15042228-15042278
3R:19498349-19498399
3L:15042249-15042328
3R:13428455-13428505
3L:22678988-22679038